Guidelines on Obtaining Ideal Performance for Parallel Analysis of Molecular Dynamics Trajectories with Python on HPC Resources

Mahzad Khoshlessan[a], Ioannis Paraskevakos[c], Geoffrey C. Fox[d], Shantenu Jha[c], Oliver Beckstein[a,b,*]

[a] *Department of Physics, Arizona State University, Tempe, AZ 85281, USA*
[b] *Center for Biological Physics, Arizona State University, Tempe, AZ 85281, USA*
[c] *Department of Electrical & Computer Engineering, Rutgers University, Piscataway, NJ 08854, USA*
[d] *Digital Science Center, Indiana University, Bloomington, IN 47405*

Abstract

In the biomolecular simulation community, analyzing trajectories from molecular dynamics (MD) simulations is becoming a bottleneck in the scientific workflow. While the trajectory generation has seen large gains in performance by leveraging modern high performance computing (HPC) resources, acceleration of the analysis step has been lagging behind. Although many typical analysis tasks can be parallelized with a Map-Reduce approach we found that obtaining near linear strong scaling performance can be very difficult in practice because of *straggler* processes whereby a few processes are much slower than the typical processes. Using the widely used *MDAnalysis* Python package as an example, we investigated a single program multiple data (SPMD) execution model where each process executes the same program to parallelize the Root Mean Square Distance (RMSD) algorithm using a Map-Reduce approach. We employed the Python language, which is widely used in the biomolecular simulation field, and focus on MPI-based implementations. Stragglers were less prevalent for compute-bound workloads (as measured by the ratio of compute

*Corresponding author
Email addresses: mkhoshle@asu.edu (Mahzad Khoshlessan), i.paraskev@rutgers.edu (Ioannis Paraskevakos), gcf@indiana.edu (Geoffrey C. Fox), shantenu.jha@rutgers.edu (Shantenu Jha), oliver.beckstein@asu.edu (Oliver Beckstein)

to I/O time and the ratio of compute to communication time). We found that stragglers were primarily caused by either excessive MPI communication costs or excessive time to open the single shared trajectory file whereas both the computation and the ingestion of data exhibited close to ideal strong scaling behavior. We improved performance by (1) reducing the communication cost with the *Global Arrays* (GA) toolkit and (2) testing two different approaches to improve file access. The first approach is splitting the trajectory into as many trajectory segments as number of processes ("subfiling"). The second approach is through MPI-based approach using Parallel HDF5 where we examine the performance through independent I/O. Applying these strategies, we obtained near ideal scaling and performance up to 384 cores. We provide insights, guidelines, and strategies to the biomolecular simulation community on how to take advantage of the available HPC resources to gain good performance.

Keywords: Python, MPI, HPC, MDAnalysis, Global Array, MPI I/O, HDF5, Straggler, Molecular Dynamics, Big Data, Trajectory Analysis
2010 MSC: 00-01, 99-00

---

1. Introduction

Molecular dynamics (MD) simulations are a powerful method to generate new insights into the function of biomolecules [1–5]. These simulations produce trajectories, timeseries of atomic coordinates, that now routinely include millions of timesteps and can be Terabytes in size. These trajectories need to be analyzed using statistical mechanics approaches [6] but because of the increasing size of data, trajectory analysis is becoming a bottleneck in typical biomolecular simulation scientific workflows [7]. Many data analysis tools and libraries have been developed to extract the desired information from the output trajectories from MD simulations [8–21] but few can efficiently use modern High Performance Computing (HPC) resources to accelerate the analysis stage. MD trajectory analysis primarily requires *reading* of data from the file system; the processed output data are typically neglibible in size compared to the

input data and therefore we exclusively investigate the reading aspects of trajectory I/O (i.e., the "I"). We focus on the *MDAnalysis* package [16, 17], which is an open-source object-oriented Python library for structural and temporal analysis of molecular dynamics simulation trajectories and individual protein structures. Although *MDAnalysis* accelerates selected algorithms with OpenMP, distributed parallel trajectory analysis has not yet been implemented in the library. Here we discuss the challenges and lessons-learned for making parallel analysis on HPC resources feasible in a general purpose trajectory analysis library such as *MDAnalysis*.

Previously, we used a parallel map-reduce approach to study the performance of calculating the minimal root mean squared distance (RMSD) of the positions of a subset of atoms to a reference conformation under optimization of rigid body translations and rotations [22, 23], also known as "RMSD fitting" [6, 24]. We investigated two parallel implementations, one using *Dask* [25] and one using *mpi4py* [26, 27]. For both *Dask* and MPI, we found that our benchmark only showed good strong scaling within a single node. Beyond a single node performance dropped due to *straggler* tasks, a subset of processes that were significantly slower than the typical execution time of all tasks; the total execution time became dominated by stragglers and overall performance decreased. Stragglers are known to significantly impede job completion time [28] and have a high impact on performance and energy consumption on big data systems [29]. In the present study, we analyzed the MPI case in more detail to better understand the origin of stragglers with the goal to find simple and robust parallelization approaches to speed up parallel post-processing of MD trajectories in the *MDAnalysis* library.

As in our previous study we selected the commonly used RMSD algorithm implemented in *MDAnalysis* as a typical use case. We used the single program multiple data (SPMD) paradigm to parallelize this algorithm on three different HPC resources (XSEDE's *SDSC Comet*, *LSU SuperMic*, and *PSC Bridges*). With SPMD, each process executes essentially the same operations on different parts of the data. The three clusters differ in their architecture except that

each uses a shared Lustre parallel file system. We used Python, a machine-independent, byte-code interpreted, object-oriented programming (OOP) language, which is well-established in the biomolecular simulation and HPC parallel communities [26, 30]. We show that there are two important performance parameters, $\overline{t_{\mathrm{comp}}}/\overline{t_{\mathrm{I/O}}}$ and $\overline{t_{\mathrm{comp}}}/\overline{t_{\mathrm{comm}}}$ which are the ratio of computation to the I/O load, as measured by the time spent on computation versus the time spent on reading data from the file system, and the ratio of computation to communication load respectively, that determine whether we observe stragglers. If the problem was compute-bound ($\overline{t_{\mathrm{comp}}}/\overline{t_{\mathrm{I/O}}} \gg 1$), the algorithm scaled very well without modifications, but otherwise strong scaling performance beyond a single node was poor. For algorithms with small $\overline{t_{\mathrm{comp}}}/\overline{t_{\mathrm{I/O}}}$, we needed to come up with strategies to improve scaling and overcome problems with stragglers. To a lesser degree, $\overline{t_{\mathrm{comp}}}/\overline{t_{\mathrm{comm}}}$ also determined scaling and performance, with $\overline{t_{\mathrm{comp}}}/\overline{t_{\mathrm{comm}}} < 1$ predicting poor performance. We show that communication and reading I/O are the two main scalability bottlenecks.

Taking advantage of Global Arrays [30, 31], we were able to reduce communication cost noticeably. However, our data showed that the initial opening of the trajectory file became the cause for stragglers beyond a single node. We examined two different approaches to mitigate I/O bottlenecks: subfiling (trajectory file splitting) with Global Arrays for communications and MPI parallel I/O with MPI for communications. Both approaches significantly improved the performance and lead to near ideal scaling.

Documentation and benchmark codes are made available in the code repository `https://github.com/hpcanalytics/supplement-hpc-py-parallel-mdanalysis` under the GNU General Public License v3.0 (code) and the Creative Commons Attribution-ShareAlike (documentation). These materials should enable users to recreate the computational environment on the tested XSEDE HPC resources (*SDSC Comet*, *PSC Bridges*, *LSU SuperMIC*), prepare data files, and run the computational experiments.

2. Background & Related Work

2.1. Stragglers

We found that straightforward implementation of simple parallelization with a Map-Reduce scheme in Python failed to scale beyond a single compute node [22] because a few tasks (MPI-ranks or Dask [25] processes) took much longer than the typical task and so limited the overall performance. However, the cause for these *straggler* tasks remained obscure. Here, we perform a detailed study of the straggler problem (also called a long tail phenomenon) and address solutions to overcome it. Long tail phenomena, whereby a small proportion of tasks significantly impede job completion time, are a challenge for improving performance [28] on HPC resources. It is a known problem in other frameworks such as Google's MapReduce [32], Spark [33–35], Hadoop [32], and cloud data centers [36]. Both internal and external factors contribute to stragglers. Internal factors include heterogeneous capacity of worker nodes, and resource competition due to other tasks running on the same worker node. External factors include resource competition due to co-hosted applications, input data skew, remote input or output source being too slow, faulty hardware [32, 37], and node mis-configuration [32]. Competition over scarce resources, in particular the network bandwidth, was found to lead to stragglers in writing on Lustre file systems [38]. Garbage collection [33, 34], JVM positioning to cores [33], delays introduced while the tasks move from the scheduler to execution [35], disk I/O during shuffling, Java's just-in-time compilation [34], and output skew [34] have also been found to introduce stragglers. In addition to these reasons, stragglers on Spark have been attributed to the overall performance of workers or competition between resources [39]. Garraghan et al. [28] reported high CPU utilization, disk utilization, unhandled I/O access requests and network package loss as the most frequent reasons for stragglers on Virtualized Cloud Data-centers.

Tuning resource allocation and parallelism such as breaking the workload into many small tasks that are dynamically scheduled at runtime [40], slow

Node-Threshold [32], speculative execution [32], sampling or data distribution estimation techniques, SkewTune to avoid data imbalance [41], dynamic work rebalancing [36], blocked time analysis [42], and intelligent scheduling [43] are among a wide variety of approaches that are trying to detect and mitigate stragglers.

In the present study, we identify the cause of the straggler tasks in the context of analyzing large MD trajectories and provide solutions through which we can improve performance and scaling. Even though the proposed solutions to avoid stragglers are specifically applied to the analysis of MD trajectories in *MDAnalysis*, all of these principles and implementation techniques are potentially applicable to data analysis in other Python-based libraries.

2.2. Other packages with parallel analysis capabilities

There have been various attempts to parallelize the analysis of MD trajectories. HiMach [13] introduces scalable and flexible parallel Python framework to deal with massive MD trajectories, by combining and extending Google's MapReduce and the VMD analysis tool [10]. HiMach's runtime is responsible to parallelize and distribute Map and Reduce classes to assigned cores. HiMach uses parallel I/O for file access during map tasks and MPI_Allgather in the reduction process. HiMach, however, does not discuss parallel analysis of analysis types that cannot be implemented via MapReduce. Furthermore, HiMach is not available under an open source license, which does not allow community contributions and continuous integration with state-of-the-art methods.

Wu et. al. [44] present a scalable parallel framework for distributed-memory post-simulation data analysis. This work consists of an interface that allows a user to write analysis programs sequentially, and the machinery that ensures these programs execute in parallel automatically. The main components of the proposed framework are (1) domain decomposition that splits computational domain into blocks with specified boundary conditions, (2) HDF5 based parallel I/O (3) data exchange that communicates ghost atoms between neighbor blocks, and (4) parallel analysis implementation of a real-world analysis application.

This work does not discuss analysis methods which cannot be implemented using MapReduce and is limited to HDF5 file format.

Zazen [45] is a novel task-assignment protocol to overcome the I/O bottleneck for many I/O bound tasks. This protocol caches a copy of simulation output files on the local disks of the compute nodes of a cluster, and uses co-located data access with computation. Zazen is implemented in a parallel disk cache system and avoids the overhead associated with querying metadata servers by reading data in parallel from local disks. The approach has been used to improve the performance of HiMach [13]. It, however, advocates a specific architecture where a parallel supercomputer, which runs the simulations, immediately pushes the trajectory data to a local analysis cluster where trajectory fragments are cached on node-local disks. In the absence of such a specific workflow, one would need to stage the trajectory across nodes, and the time for data distribution is likely to reduce any gains from the parallel analysis.

VMD [10, 46] provides molecular visualization and analysis tool through algorithmic and memory efficiency improvements, vectorization of key CPU algorithms, GPU analysis and visualization algorithms, and good parallel I/O performance on supercomputers. It is one of the most advanced programs for the visualization and analysis of MD simulations. It is, however, a large monolithic program, that can only be driven through its built-in Tcl interface and thus is less well suited as a library that allows the rapid development of new algorithms or integration into workflows.

CPPTraj [18] offers multiple levels of parallelization (MPI and OpenMP) in a monolithic C++ implementation. CCPTraj allows parallel reads between frames of the same trajectory but is especially geared towards processing an ensemble of many trajectories in parallel.

pyPcazip [47] is a suite of software tools written in Python for compression and analysis of molecular dynamics (MD) simulation data. pyPcazip is MPI parallelised and is specific to PCA-based investigations of MD trajectories and supports a wide variety of trajectory file formats (based on the capabilities of the underlying mdtraj package [19]). pyPcazip can take one or many input MD

trajectory files and converts them into a highly compressed, HDF5-based, .pcz format with insignificant loss of information. However, the package does not support general purpose analysis.

In situ analysis is an approach to simultaneously execute analysis while the MD simulations are running so that I/O bottlenecks are potentially mitigated [48]. Malakar et al. studied the scalability challenges of time — and space — shared modes of analyzing large-scale MD simulations through a topology-aware mapping for simulation and analysis by using LAMMPS code.

All of the above frameworks provide tools for parallel analysis of MD trajectories. These frameworks, however, tend to fall short in providing parallelism in the context of a general and flexible library for the analysis of MD trajectories. Although straggler tasks are a common challenge arising in parallel analysis and are well-known in the data analysis community (see Section 2.1), there is, to our knowledge, little discussion about this problem in the biomolecular simulation community. Our own experience with a MapReduce approach in *MDAnalysis* [22] suggested that stragglers might be a somewhat under-appreciated problem. Therefore, in the present work we want to better understand requirements for efficient parallel analysis of MD trajectories in *MDAnalysis*, but to also provide more general guidance that could benefit developments in other libraries inside and outside of the scope of analysis of MD simulations.

3. Software Modules Used

3.1. MDAnalysis

Simulation data exist in trajectories in the form of time series of atom positions and sometimes velocities. These come in a plethora of different and idiosyncratic file formats. *MDAnalysis* [16, 17] is a widely used open source library to analyze trajectory files with an object oriented interface. The library is written in Python, with time critical code in C/Cython. *MDAnalysis* supports most file formats of simulation packages including CHARMM, Gromacs,

Amber, and NAMD and the Protein Data Bank format and enables accessing data stored in trajectories.

As a test case that is representative of a common task in the analysis of biomolecular simulation trajectories we chose the calculation of a timeseries of the minimal structural root mean square distance $\mathrm{RMSD}(t) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\delta_i(t)^2}$ (**RMSD**), where δ_i is the distance between atom i at time t and its position in a reference structure [6, 49]. The RMSD is calculated for the C_α atoms of a subset of protein residues after optimal superposition with the QCPROT algorithm [24, 50] (implemented in Cython and available through the `MDAnalysis.analysis.rms` module [17]), as outlined in Algorithm 1. In the RMSD algorithm, a translation and rigid body rotation are found that minimize the RMSD between the coordinates at time t and the reference coordinates [24]. The RMSD of this optimum superposition is reported for each time step t. The RMSD is used to show the rigidity of protein domains and more generally characterizes structural changes. The time complexity for the RMSD Algorithm 1 is $\mathcal{O}(T \times N^2)$ [24] where T is the number of frames in the trajectory and N the number of particles included in the RMSD calculation.

Algorithm 1 MPI-parallel Multi-frame RMSD Algorithm

Input: *size*: Total number of frames
ref: mobile group in the initial frame which will be considered as reference
start & stop: Starting and stopping frame index
topology & trajectory: files to read the data structure from
Output: Calculated RMSD arrays

procedure $Block\_RMSD$(topology, trajectory, ref, start, stop)
    u ← Universe(topology, trajectory) ▷ u hold all the information of the physical system
    g ← u.frames[start:stop]
    for $\forall iframe$ in g **do**
        $results[iframe] \leftarrow RMSD(g, ref)$
    end for
    return results
end procedure

MPI Init
rank ← rank ID
index ← indices of mobile atom group
xref0 ← Reference atom group's position
out ← Block_RMSD(topology, trajectory, xref0, start=start, stop=stop)

Gather(out, RMSD_data, rank_ID=0)
MPI Finalize

3.2. MPI for Python (mpi4py)

MPI for Python (*mpi4py*) is a Python wrapper written over Message Passing Interface (MPI) standard and allows any Python program to employ multiple processors [26, 27]. Python is widely used in the scientific community because it facilitates rapid development of small scripts and code prototypes as well as development of large applications and highly portable and reusable modules and libraries. Based on the efficiency tests [26, 27], the performance degradation due to using *mpi4py* is not prohibitive and the overhead introduced by *mpi4py* is far smaller than the overhead associated to the use of interpreted versus compiled languages [30]. Overheads in *mpi4py* are small compared to C code if efficient raw memory buffers are used for communication [26].

3.3. Global Arrays toolkit

The *Global Arrays* (GA) toolkit provides users with a language interface that allows them to distribute data while maintaining the type of global index space and programming syntax similar to what is available when programming on a single processor [31]. It allows manipulating physically distributed dense multi-dimensional arrays without explicitly defining communication and synchronization between processes. The underlying communication is determined by a runtime environment, which defaults to the *Communication runtime for Extreme Scale* (ComEx) [51]. ComEx uses shared memory for intra-node communication and inter-node communication employs ComEx with MPI. *Global Arrays in NumPy* (GAiN) extends GA to Python through Numpy [30]. The Global Arrays toolkit provides functions to create global arrays (`ga_create()`) and to copy data to (`ga_put()`) and from (`ga_get()`) such a global array, as well as additional functions for copying between arrays and freeing them [30]. When a global array is created (`ga_create()`) each process will create an array of the same shape and size, physically located in the local memory space of that process [31]. The GA library maintains a list of all these memory locations, which can be queried with the `ga_access()` function. Using a pointer returned by `ga_access()`, one can directly modify the data that is local to

each process. When a process tries to access a block of data the request is first decomposed into individual blocks representing the contribution to the total request from the data held locally on each process (*B. J. Palmer and J. Daily, personal communication*). The requesting process then makes individual requests to each of the remote processes.

Algorithm 2 describes the RMSD algorithm in combination with a Global Arrays. In this algorithm, we use Global Arrays instead of the message passing paradigm to investigate if we can reduce the communication cost.

Algorithm 2 MPI-parallel Multi-frame RMSD using Global Arrays

Input:*size*: Total number of frames assigned to each rank N_b
g_a: Initialized Global Arrays
xref0: mobile group in the initial frame which will be considered as reference
start & stop: that tell which block of trajectory (frames) is assigned to each rank
topology & trajectory: files to read the data structure from
Include: `Block_RMSD()` from Algorithm 1

```
bsize ← ceil(trajectory.number_frames / size)
g_a ← ga.create(ga.C_DBL, [bsize*size,2], "RMSD")
buf ← np.zeros([bsize*size,2], dtype=float)
out ← Block_RMSD(topology, trajectory, xref0, start=start, stop=stop)
ga.put(g_a, out, (start,0), (stop,2))
if rank == 0 then
    buf ← ga.get(g_a, lo=None, hi=None)
end if
```

3.4. MPI and Parallel HDF5

HDF5 is a structured self-describing hierarchical data format which is the standard mechanism for storing large quantities of numerical data in Python (`http://www.hdfgroup.org/HDF5`, [52]). Parallel HDF5 (*PHDF5*) typically sits on top of a MPI-IO layer and can use MPI-IO optimizations. In *PHDF5*, all file accesses are coordinated through the MPI library; otherwise, multiple processes would compete over accessing over the same file on disk. MPI-based applications work by launching multiple parallel instances of the Python interpreter which communicate with each other via the MPI library. *HDF5* itself handles nearly all the details involved with coordinating file access when the shared file is opened through the *mpio* driver. In addition, MPI communicator should be supplied as well and the users also need to follow some constraints for data consistency [52].

MPI has two flavors of operation: collective, which means that all processes have to participate in the same order, and independent, which means each process can perform the operation or not and the order also does not matter [52]. With *PHDF5*, modifications to file metadata must be done collectively and although all processes perform the same task, they do not wait until the others catch up [52]. Other tasks and any type of data operations can be performed independently by processes. In the present study, we use independent operations.

4. Benchmark Environment

Our benchmark environment consisted of three different XSEDE HPC resources (described in section 4.1), the software stack used (section 4.2), which had to be compiled for each resource, and the common test data set (section 4.3).

4.1. HPC Resources

The computational experiments were executed on standard compute nodes of three XSEDE supercomputers, *SDSC Comet*, *PSC Bridges*, and *LSU SuperMIC* (Table 1). *SDSC Comet* is a 2.7 PFlop/s cluster with 6,400 compute nodes in total. It is optimized for running a large number of medium-size calculations (up to 1,024 cores) to support the most prevalent type of calculation on XSEDE resources. *PSC Bridges* is a 1.35 PFlop/s cluster with different types of computational nodes, including 16 GPU nodes, 8 large memory and 2 extreme memory nodes, and 752 regular nodes. It was designed to flexibly support both traditional (medium scale calculations) and non-traditional (data analytics) HPC uses. *LSU SuperMIC* offers 360 standard compute nodes with a peak performance of 557 TFlop/s.

4.2. Software

Table 2 lists the tools and libraries that were required for our computational experiments. Many domain specific packages are not available in the standard software installation on super-computers. We therefore had to compile them,

Name	Nodes	Number of Nodes	CPUs	RAM	Network Topology	Scheduler and Resource Manager
SDSC Comet	Compute	6400	2 Intel Xeon (E5-2680v3) 12 cores/CPU, 2.5 GHz	128 GB DDR4 DRAM	56 Gbps IB	SLURM
PSC Bridges	RSM	752	2 Intel Haswell (E5-2695 v3) 14 cores/CPU, 2.3 GHz	128 GB, DDR4-2133Mhz	12.37 Gbps OPA	SLURM
SuperMIC	Standard	360	2 Intel Ivy Bridge (E5-2680) 10 cores/CPU, 2.8 GHz	64 GB, DDR3-1866Mhz	56 Gbps IB	PBS

Table 1: Configuration of the HPC resources that were benchmarked. Only a subset of the total available nodes were used. IB: InfiniBand; OPA: Omni-Path Architecture.

which in some cases required substantial effort due to non-standard building and installation procedures or lack of good documentation. Because this is a common problem that hinders reproducibility we provide detailed version information, notes on the installation process, as well as comments on the ease of installation and the quality of the documentation in Table 2. For the MPI implementation we used Open MPI release 1.10.7 (`https://www.open-mpi.org/`) consistently everywhere. Detailed instructions to create the computing environments together with the benchmarking code can be found in the GitHub repository. Carefully setting up the same software stack on the three different supercomputers allowed us to clearly demonstrate the reproducibility of our results and showed that our findings were not dependent on machine specifics.

4.3. Data set

The test system contained the protein adenylate kinase with 214 amino acid residues and 3341 atoms in total [53] and the topology information (atoms types and bonds) was stored in a file in CHARMM PSF format. The test trajectory was created by concatenating 600 copies of a molecular dynamics trajectory with 4,187 time frames (saved every 240 ps for a total simulated time of 1.004 μs) in CHARMM DCD format [54] and converting to Gromacs XTC format trajectory, as described for the "600x" trajectory in Khoshlessan et al. [22]. The trajectory has a file size of about 30 GB and contains 2,512,200 frames (corresponding to 602.4 μs simulated time). The file size is relatively small because water molecules that were also part of the original MD simulations were stripped to reduce the original file size by a factor of about 10; such preprocessing is a common approach if one is only interested in the protein behavior. Thus, the trajectory represents a small to medium system size in the number of atoms

Package	Version	Description	Ease of Installation	Documentation	Installation	Dependencies
GCC	4.9.4	GNU Compiler Collection	0	++	via configuration files, environment or command line options, minimal configuration is required	–
Open MPI	1.10.7	MPI Implementation	0	++	via configuration files, environment or command line options, minimal configuration is required	–
Global Arrays	5.6.1	Global Arrays	–	+	via configuration files, environment or command line options, several optional configuration settings available	MPI 1.x/2.x/3.x implementation like Open MPI built with shared/dynamic libraries, GCC
MPI4py	3.0.0	MPI for Python	+	++	Conda Installation	Python 2.7 or above, MPI 1.x/2.x/3.x implementation like Open MPI built with shared/dynamic libraries, Cython
GA4py	1.0	Global Arrays for Python	0	0	Python Setuptools	Python 2 only, MPI 1.x/2.x/3.x implementation like Open MPI built with shared/dynamic libraries, Cython, MPI4py, Numpy
PHDF5	1.10.1	Parallel HDF5	–	++	via configuration files, environment or command line options, several optional configuration settings available	MPI 1.x/2.x/3.x implementation like Open MPI GNU, MPIF90, MPICC, MPICXX
H5py	2.7.1	Pythonic wrapper around the HDF5	+	++	Conda Installation	Python 2.7, or above, PHDF5, Cython
MDAnalysis	0.17.0	Python library to analyze trajectories from MD simulations	+	++	Conda Installation	Python >=2.7, Cython, GNU, Numpy

Table 2: Detailed comparison on the dependencies and installation of different software packages used in the present study. Software was built from source or obtained via a package manager and installed on the multi-user HPC systems in Table 1. Evaluation of ease of installation and documentation uses a subjective scale with "++" (excellent), "+" (good), "0" (average), and "−" (difficult/lacking) and reflects the experience of a typical domain scientist at the graduate/post-graduate level in a discipline such as computational biophysics or chemistry.

and coordinates that have to be loaded into memory for each time frame. The XTC format is a format with lossy compression, which also contributes to the compact file size and which trades lower I/O demands for higher CPU demands during decompression and therefore performed well in our previous study [22]. Although, 2,512,200 frames represents a long simulation for current standards, such trajectories will become increasingly common due to the use of special hardware [55] and GPU-acceleration [56, 57].

5. Methods

5.1. Timing observables

We evaluate MPI performance based on the RMSD algorithm 1. The notation for our models is summarized in Table 3. We abbreviate the timings in the following as variables $t_{\mathrm{L_n}}$ where L_n refers to the line number in algorithm 1.

We directly measured inside our code (in the function `block_rmsd()`) the "I/O" time for ingesting the data from the file system into memory, ($t_{\mathrm{I/O}}^{\mathrm{frame}} = t_{\mathrm{L4}}$) and the "compute" time per trajectory frame to perform the computation ($t_{\mathrm{comp}}^{\mathrm{frame}} = t_{\mathrm{L5}}$). The total I/O time for a rank $t_{\mathrm{I/O}} = \sum_{\mathrm{frame}=1}^{N_{\mathrm{b}}} t_{\mathrm{I/O}}^{\mathrm{frame}}$ is the sum over all I/O times for all the N_{frames} frames assigned to the rank; similarly, the total compute time for a rank is $t_{\mathrm{comp}} = \sum_{\mathrm{frame}=1}^{N_{\mathrm{b}}} t_{\mathrm{comp}}^{\mathrm{frame}}$. The time delay between the end of the last iteration and exiting the `for` loop is $t_{\mathrm{end\_loop}} = t_{\mathrm{L6}} + t_{\mathrm{L7}}$. The time $t_{\mathrm{opening\_trajectory}} = t_{\mathrm{L2}} + t_{\mathrm{L3}}$ measures the problem setup, which includes data structure initialization and opening of topology and trajectory files. $t_{\mathrm{Communication\_MPI}} = t_{\mathrm{L16}}$ is the time to gather ("reduce") all data from all processor ranks to rank zero. The total time (for all frames) spent in `block_rmsd()` is $t_{\mathrm{RMSD}} = \sum_{i=1}^{8} t_{\mathrm{Li}}$. There are parts of the code in `block_rmsd()` that are not covered by the detailed timing information of t_{comp} and $t_{\mathrm{I/O}}$.

Unaccounted time is considered as "overhead". We define $t_{\mathrm{Overhead1}}$ and $t_{\mathrm{Overhead2}}$ as the overhead of the calculations. They should ideally be very small (see Table 3 for the definition). Finally, the total time to completion of a single process, when utilizing N cores for the execution of the overall experiment, is t_N, as a result $t_{\mathrm{RMSD}} + t_{\mathrm{comm}} \equiv t_N$.

5.2. Performance Measurement

We also recorded the total time to solution $t_{\mathrm{total}}(N)$ with N MPI processes on N cores (which is effectively $t_{\mathrm{total}}(N) \approx \max(t_N)$). Strong scaling was quantified by calculating the speed-up relative to performance on a single core

$$S = \frac{t_{\mathrm{total}}(N)}{t_{\mathrm{total}}(1)} \tag{1}$$

and efficiency

$$E = \frac{S}{N}. \tag{2}$$

All averages are defined as

$$\overline{t_{comp}} = \frac{1}{N}\sum_{\text{rank}=1}^{N} t_{\text{comp}} = \frac{1}{N}\sum_{\text{rank}=1}^{N}\sum_{\text{frame}=1}^{N_\text{b}} t_{\text{comp}}^{\text{frame}}, \tag{3}$$

$$\overline{t_{\text{comm}}} = \frac{1}{N}\sum_{\text{rank}=1}^{N} t_{\text{communication}} \tag{4}$$

$$\overline{t_{\text{I/O}}} = \frac{1}{N}\sum_{\text{rank}=1}^{N} t_{\text{I/O}} = \frac{1}{N}\sum_{\text{rank}=1}^{N}\sum_{\text{frame}=1}^{N_\text{b}} t_{\text{I/O}}^{\text{frame}} \tag{5}$$

Additionally, we introduced two performance parameters that we will show are indicative of the occurrence of stragglers. We define the ratio of compute time to I/O time as $t_{\text{comp}}/t_{\text{I/O}}$. In the present study, we calculated this ratio using the serial version of our algorithms because compute time per frame and I/O time per frame should ideally be the same for different runs using different processes, as we observed previously [22]. The ratio of compute to communication time is defined by the ratio of total compute time per rank averaged across all ranks to the total communication time per rank averaged across all ranks

$$\frac{\overline{t_{\text{comp}}}}{\overline{t_{\text{comm}}}} \tag{6}$$

Table 3 offers a summary of all notation.

All code used in the present study is available in our Github repository.

6. Computational Experiments

6.1. RMSD Benchmark

The RMSD algorithm represents a task whose computational load is smaller than the I/O load per frame (typical values $t_{\text{comp}}^{\text{frame}} = 0.09$ ms, $t_{\text{IO}}^{\text{frame}} = 0.3$ ms, thus $\overline{t_{\text{comp}}}/\overline{t_{\text{I/O}}} \approx 0.3$). We showed in [22] that the RMSD task only scaled well up on a single compute node on *SDSC Comet* and *TACC Stampede*, using either Dask or MPI.

Item	Definition
N_b	$N^\mathrm{total}_\mathrm{frames}/N$
$t_\mathrm{end\_loop}$	$t_\mathrm{L6} + t_\mathrm{L7}$
$t_\mathrm{opening\_trajectory}$	$t_\mathrm{L2} + t_\mathrm{L3}$
t_comp	$\sum_\mathrm{frame=1}^{N_\mathrm{b}} t^\mathrm{frame}_\mathrm{comp}$
$t_\mathrm{I/O}$	$\sum_\mathrm{frame=1}^{N_\mathrm{b}} t^\mathrm{frame}_\mathrm{I/O}$
$t_\mathrm{all\_frame}$	$t_\mathrm{L4} + t_\mathrm{L5} + t_\mathrm{L6}$
t_RMSD	$t_\mathrm{L1} + ... + t_\mathrm{L8}$
$t_{\mathrm{Communication}_\mathrm{MPI}}$	t_L16
$t_{\mathrm{Communication}_\mathrm{GA}}$	$t_\mathrm{L5} + t_\mathrm{L6} + t_\mathrm{L7} + t_\mathrm{L8}$
$t_\mathrm{Overhead1}$	$t_\mathrm{all\_frame} - t_\mathrm{I/O\_final} - t_\mathrm{comp\_final} - t_\mathrm{end\_loop}$
$t_\mathrm{Overhead2}$	$t_\mathrm{RMSD} - t_\mathrm{all\_frame} - t_\mathrm{opening\_trajectory}$
t_N	$t_\mathrm{RMSD} + t_\mathrm{Communication}$
$\overline{t_\mathrm{comp}}$	$\frac{1}{N}\sum_\mathrm{rank=1}^{N}\sum_\mathrm{frame=1}^{N_\mathrm{b}} t^\mathrm{frame}_\mathrm{comp}$
$\overline{t_\mathrm{comm}}$	$\frac{1}{N}\sum_\mathrm{rank=1}^{N} t_\mathrm{communication}$
$\overline{t_\mathrm{I/O}}$	$\frac{1}{N}\sum_\mathrm{rank=1}^{N}\sum_\mathrm{frame=1}^{N_\mathrm{b}} t^\mathrm{frame}_\mathrm{I/O}$
t_total	$\max t_N$

Table 3: Summary of the notation of our performance modeling. Relevant probes in the codes are taken and stored, which we will abbreviate in here as t_Ln where Ln refers to the line number in the corresponding algorithm. $t_\mathrm{Communication\_MPI}$ and $t_\mathrm{Communication\_GA}$ are both referred to $t_\mathrm{Communication}$ in the text. All the timings on the first row are per rank.

We focus on the MPI implementation (via *mpi4py* [26, 27]) to better understand the cause for the lack of scaling across nodes. As in our previous work, we also observed poor strong scaling performance (Figures 1a, 1b and A.10a and A.10b) beyond a single node. A more detailed analysis showed that the RMSD computation, and to a lesser degree the I/O, considered on their own scaled well beyond 50 cores (yellow and blue lines in Figure 1c). But, communication (red line in Figure 1c) and the initial file opening (gray line in Figure 1c) started to dominate beyond 50 cores. Communication cost and initial time for opening the trajectory were distributed unevenly across MPI ranks, as shown in Figure 1d. The ranks that took much longer to complete than the typical execution time of the other ranks were the stragglers that hurt performance.

Identification of Scalability Bottlenecks

In the example shown in Figure 1d, 62 ranks out of 72 took about 60 s (the stragglers) whereas the remaining ranks only took about 20 s. The detailed

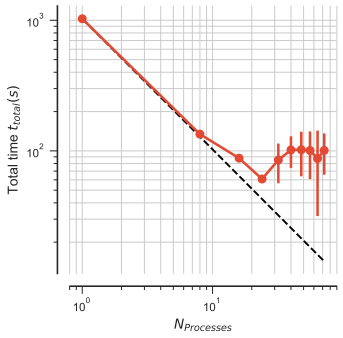
breakdown of the time spent on each rank (Figure 1d) showed that the computation, t_{comp}, was relatively constant across ranks. The time spent on reading data from the shared trajectory file on Lustre into memory, $t_{\mathrm{I/O}}$, showed variability across different ranks. The stragglers, however, appeared to be defined by occasionally much larger *communication* times, t_{comm} (line 16 in Algorithm 1), which were on the order of 30 s, and by larger times to initially open the trajectory (line 2 in Algorithm 1). t_{comm} varied across different ranks and was barely measurable for a few of them. Although the data in Figure 1d represent one run and in other instances different number of ranks were stragglers, the overall hypothesis was confirmed by the averages over five independent repeats and all ranks (Figure 1c). This initial analysis indicated that communication was a major issue that prevented good scaling beyond a single node but the problems related to file I/O also played an important role in limiting performance.

Influence of Hardware

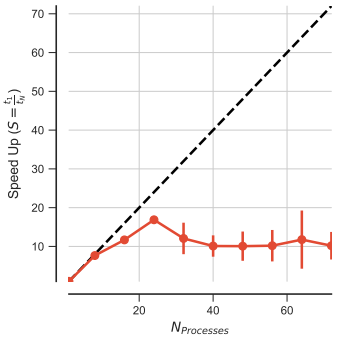
We ran the same benchmarks on multiple HPC systems (XSEDE's *PSC Bridges* (Fig. A.10) and *LSU SuperMIC* (Fig. A.11)), and observed the occurrence of stragglers, in a manner very similar to the results described for *SDSC Comet*. There was no clear pattern in which certain MPI ranks would always be a straggler, neither could we trace stragglers to specific cores or nodes. Therefore, the phenomenon of stragglers in the RMSD case was reproducible on different clusters and thus appeared to be independent from the underlying hardware.

6.2. Effect of average compute over average IO ratio on Performance

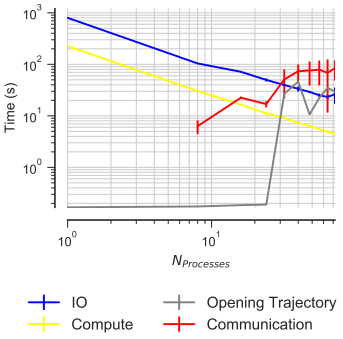
The results in section 6.1 indicated communication and I/O to be two important factors that appeared to correlate with stragglers. In order to better characterize the RMSD task, we computed the ratio between the time to complete the computation and the time spent on I/O, $\overline{t_{\mathrm{comp}}}/\overline{t_{\mathrm{I/O}}} = 0.3$, and found



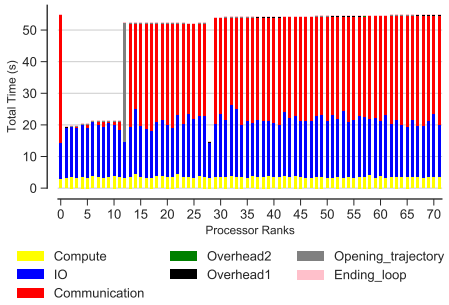
(a) Scaling total (five repeats)



(b) Speed-up (five repeats)



(c) Scaling for different components (five repeats)



(d) Time comparison on different parts of the calculations per MPI rank (example)

Figure 1: Performance of the RMSD task (I/O-bound with $\overline{t_{\mathrm{comp}}}/\overline{t_{\mathrm{I/O}}} \approx 0.3$) with MPI on *SDSC Comet*. Results were communicated back to rank 0. Five independent repeats were performed to collect statistics. (a-c) The error bars show standard deviation with respect to mean. (d) Compute t_{comp}, IO $t_{\mathrm{I/O}}$, communication t_{comm}, ending the for loop $t_{\mathrm{end\_loop}}$, opening the trajectory $t_{\mathrm{opening\_trajectory}}$, and overheads $t_{\mathrm{overhead1}}$, $t_{\mathrm{overhead2}}$ per MPI rank (see Table 3 for the definition). These are data from one run of the five repeats. MPI ranks 0, 12–27 and 29–72 are stragglers. **Note:** In serial, there is no communication.

that this task was I/O bound, i.e.,

$$\frac{\overline{t_{\mathrm{comp}}}}{\overline{t_{\mathrm{I/O}}}} \ll 1.$$

For such a I/O-bound task we were not able to achieve good scaling beyond a single node. We hypothesized that decreasing the relative I/O load with respect to the compute load would also reduce the impact of stragglers by interleaving I/O with longer periods of computation and thus reducing the impact of I/O contention. We therefore set out to measure compute bound tasks, i.e. ones with

$$\frac{\overline{t_{\mathrm{comp}}}}{\overline{t_{\mathrm{I/O}}}} \gg 1.$$

To measure the effect of the $\overline{t_{\mathrm{comp}}}/\overline{t_{\mathrm{I/O}}}$ ratio on performance but leaving other parameters the same, we artificially increased the computational load by repeating the same RMSD calculation (line 10, algorithm 1) 40, 70 and 100 times in a loop, resulting in forty-fold ("40×"), seventy-fold ("70×"), and one hundred-fold ("100×") load increases.

6.2.1. Effect of Increased Compute workload for RMSD Task

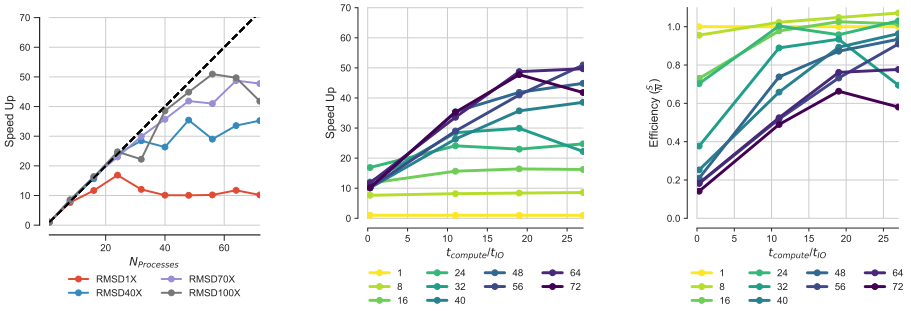
The increased computational workloads corresponded to $\overline{t_{\mathrm{comp}}}/\overline{t_{\mathrm{I/O}}}$ ratios of 11, 19, 27 respectively as shown in Table 4. Theoretically, the $\overline{t_{\mathrm{comp}}}/\overline{t_{\mathrm{I/O}}}$ ratio for the higher workloads (with factor X) should be X times the $\overline{t_{\mathrm{comp}}}/\overline{t_{\mathrm{I/O}}}$ for the 1× workload because the average I/O workload of each rank is $N_{\mathrm{b}} \times t_{\mathrm{I/O}}$ (independent of X) while the workload for the computation is $X \times N_{\mathrm{b}} \times t_{\mathrm{comp}}$. Therefore the ratio at higher X should be X times the $\overline{t_{\mathrm{comp}}}/\overline{t_{\mathrm{I/O}}}$ at 1×. The data in Table 4 agree with this theoretical analysis.

We performed the experiments with increased workload to measure the effect of the $\overline{t_{\mathrm{comp}}}/\overline{t_{\mathrm{I/O}}}$ ratio on performance (Figure 2). As the $\overline{t_{\mathrm{comp}}}/\overline{t_{\mathrm{I/O}}}$ ratio increased, speed-up and performance improved, and showed overall better scaling than the I/O-bound workload, i.e. 1× RMSD (Figure 2a). The RMSD calcu-

Workload	$t_{\mathbf{comp}}$	$t_{\mathbf{I/O}}$	$\overline{t_{\mathbf{comp}}}/\overline{t_{\mathbf{I/O}}}$	
			measured	theoretical
1×	226	791	0.29	
40×	8655	791	11	11
70×	15148	791	19	20
100×	21639	791	27	29

Table 4: Change in $\overline{t_{\mathrm{comp}}}/\overline{t_{\mathrm{I/O}}}$ ratio with change in the RMSD workload. The RMSD workload was artificially increased in order to examine the effect of compute to I/O ratio on the performance. The reported compute and I/O time were calculated based on the serial version using one core and used to calculate the measured $\overline{t_{\mathrm{comp}}}/\overline{t_{\mathrm{I/O}}}$. The theoretical $\overline{t_{\mathrm{comp}}}/\overline{t_{\mathrm{I/O}}}$ (see text) is provided for comparison.

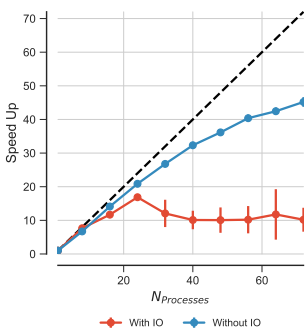
lation consistently scaled well up to larger numbers of cores ($N = 56$ for 70× RMSD) for higher $\overline{t_{\mathrm{comp}}}/\overline{t_{\mathrm{I/O}}}$ ratios. Figures 2b and 2c show that speed-up and efficiency approach their ideal value for each processor count with increasing $\overline{t_{\mathrm{comp}}}/\overline{t_{\mathrm{I/O}}}$ ratio.



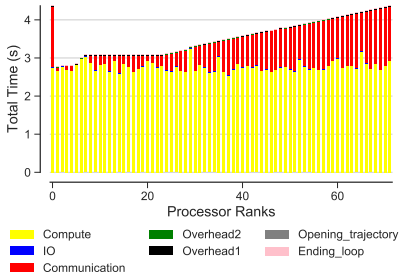
(a) Speed-Up (b) Speed-Up (c) Efficiency

Figure 2: Effect of $\overline{t_{\mathrm{comp}}}/\overline{t_{\mathrm{I/O}}}$ ratio on performance of the RMSD task with MPI performed on *SDSC Comet*. We tested performance for $\overline{t_{\mathrm{comp}}}/\overline{t_{\mathrm{I/O}}}$ ratios of 0.3, 11, 19, 27. which correspond to 1× RMSD, 40× RMSD, 70× RMSD, and 100× RMSD respectively. (a) Effect of $\overline{t_{\mathrm{comp}}}/\overline{t_{\mathrm{I/O}}}$ on the Speed-Up (b) Change in the Speed-Up with respect to $\overline{t_{\mathrm{comp}}}/\overline{t_{\mathrm{I/O}}}$ for different processor counts (c) Change in the efficiency with respect to $\overline{t_{\mathrm{comp}}}/\overline{t_{\mathrm{I/O}}}$ for different processor counts

Even for moderately compute-bound workloads, such as the 40× and 70× RMSD tasks, increasing the computational workload over I/O reduced the impact of stragglers even though they still contributed to large variations in timing across different ranks and thus irregular scaling.



(a) Speed up comparison



(b) Time comparison on different parts of the calculations per MPI rank when I/O is removed

Figure 3: Comparison on the performance of RMSD task with MPI with I/O and without I/O ($\overline{t_{\mathrm{comp}}}/\overline{t_{\mathrm{I/O}}} \approx 0.3$) on SDSC Comet. Five repeats were performed to collect statistics. (a) Speed-up. The error bars show standard deviation with respect to mean. (b) Compute t_{comp}, IO $t_{\mathrm{I/O}}$, communication t_{comm}, ending the for loop $t_{\mathrm{end\_loop}}$, opening the trajectory $t_{\mathrm{opening\_trajectory}}$, and overheads $t_{\mathrm{overhead1}}$, $t_{\mathrm{overhead2}}$ per MPI rank (See Table 3 for the definition).

6.2.2. I/O leads to stragglers

In order to study an extreme case of a compute-bound task, we eliminated all I/O from the RMSD task by generating artificial trajectory data randomly in memory of the same size as would have been obtained from the trajectory; the time for the data generation was excluded and no file access was necessary. Without any I/O, performance improved markedly (Figure 3), with reasonable scaling up to 72 cores (3 nodes). No stragglers were observed although an increase in communication time prevented ideal scaling performance. Although in practice I/O cannot be avoided, this experiment demonstrated that accessing the trajectory file on the Lustre file system is at least one cause for the observed stragglers.

6.3. Reducing Communication Cost: Application of Global Arrays

As seen in Figure 1d for small $\overline{t_{\mathrm{comp}}}/\overline{t_{\mathrm{I/O}}}$, communication acted as a scalability bottleneck. When the processes communicated result arrays back to the master process (rank 0), some processes took much longer as compared to others. We therefore investigated strategies to reduce communication cost.

We used Global Arrays (GA) [30, 31] instead of collective communication in MPI and examined the change in the performance. In GA, we define one large

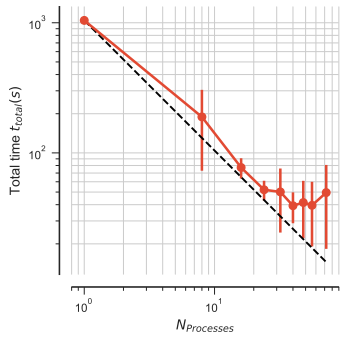
RMSD array called *global array* and each MPI rank updates its associated block in the global RMSD array using `ga_put()`. At the end, when all the processes exit `block_rmsd()` function and update their local block in the global array, rank 0 will access the whole global array using `ga_access()`. In GA, the time for communication is $t_{\text{ga\_put()}} + t_{\text{ga\_access()}}$.

Using Global Arrays improved the strong scaling performance (Figures 4a and 4b) by reducing the communication time. Nevertheless, the remaining variation in the trajectory I/O part of the calculation and in particular the initial opening of the trajectory prevented ideal scaling (Figure 4c). Figure 4d shows that stragglers were primarily due to the fact that all ranks had to open the same trajectory file at the beginning of the execution. In this case, these slow processes took about 50 s, which was slower than the mean execution time of all other ranks of 17 s. Trajectory opening was already problematic in the initial test (Figure 1c), which was still dominated by the communication cost. By substantially reducing communication cost, the simultaneous trajectory opening by multiple ranks emerged as the next dominant cause for stragglers. The improvement in performance can be attributed to the mitigation in the interference of MPI traffic with IO traffic as also studied in [58].

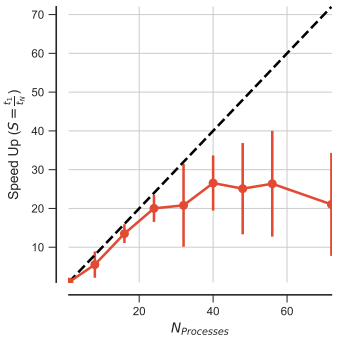
Motivated by the results in this section, we investigated the influence of the ratio of compute to communication costs ($\overline{t_{\text{comp}}}/\overline{t_{\text{comm}}}$) on performance in Appendix B. We found evidence to support the hypothesis that a larger ratio was beneficial, provided I/O costs could also be reduced, as discussed in the next section.

6.4. Reducing I/O Cost

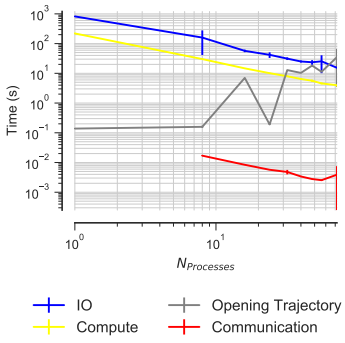
In order to improve performance we needed to employ strategies to avoid the competition over file access across different ranks when the $\overline{t_{\text{comp}}}/\overline{t_{\text{I/O}}}$ ratio was small. To this end, we experimented with two different ways for reducing the I/O cost: 1) splitting the trajectory file into as many segments as the number of processes, thus using file-per-process access, and 2) using the HDF5 file format together with MPI-IO parallel reads instead of the XTC trajectory format. We



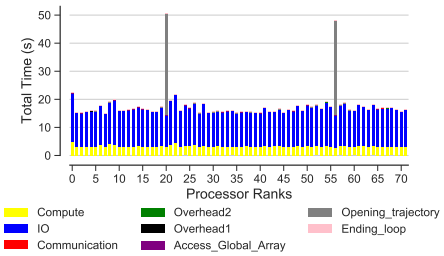
(a) Scaling total



(b) Speed-up



(c) Scaling for different components



(d) Time comparison on different parts of the calculations per MPI rank

Figure 4: Performance of the RMSD task with MPI using Global Arrays ($\overline{t_{\mathrm{comp}}}/\overline{t_{\mathrm{I/O}}} \approx 0.3$) on SDSC Comet. All ranks update the Global Arrays (`ga_put()`) and rank 0 accesses the whole RMSD array through the global memory address (`ga_get()`). Five repeats were performed to collect statistics. (a-c) The error bars show standard deviation with respect to mean. (d) Compute t_{comp}, IO $t_{\mathrm{I/O}}$, communication t_{comm}, access to the whole Global Arrays by rank 0, $t_{\mathrm{Access\_Global\_Array}}$, ending the for loop $t_{\mathrm{end\_loop}}$, opening the trajectory $t_{\mathrm{opening\_trajectory}}$, and overheads $t_{\mathrm{overhead1}}$, $t_{\mathrm{overhead2}}$ per MPI rank (See Table 3 for the definition). This is typical data from one run of the 5 repeats. MPI ranks 20 and 56 are stragglers, i.e., their total time far exceeds the mean of the all ranks. **Note:** In serial, there is no communication.

discuss these two approaches and their performance improvements in detail in the following sections.

6.4.1. Splitting the Trajectories ("subfiling")

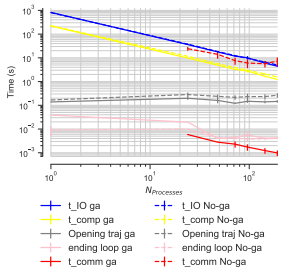
Subfiling is a mechanism previously used for splitting a large multi-dimensional global array to a number of smaller subarrays in which each smaller array is saved in a separate file. Subfiling reduces the file system control over-

head by decreasing the number of processes concurrently accessing a shared file [59, 60]. Because subfiling is known to improve programming flexibility and performance of parallel shared-file I/O, we investigated splitting our trajectory file into as many trajectory segments as the number of processes. The trajectory file was split into N segments, one for each process, with each segment having N_b frames. This way, each process would access its own trajectory segment file without competing for file accesses.

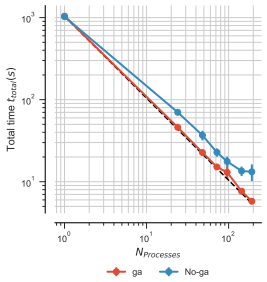
Performance with MPI communication. We ran a benchmark up to 8 nodes (192 cores) and observed rather better scaling behavior with efficiencies above 0.6 (Figure 5b and 5c) with the delay time for stragglers reduced from 65 s to about 10 s for 72 processes. However, scaling is still far from ideal due to communication (using MPI). Although the delay due to communication was much smaller as compared to parallel RMSD with shared-file I/O (compare Figure 5d ($t_{\mathrm{comm}}^{\mathrm{Straggler}} \gg t_{\mathrm{comp}} + t_{\mathrm{I/O}}$) to Figure 1d ($t_{\mathrm{comm}}^{\mathrm{Straggler}} \approx t_{\mathrm{comp}} + t_{\mathrm{I/O}}$)), it is still delaying several processes and resulted in longer job completion times (Figure 5d). These delayed tasks impacted performance so that speed-up remained far from ideal (Figure 5c).

Performance using Global Arrays. In Section 6.3 we showed that Global Arrays substantially reduced the communication cost. We wanted to quantify the performance when splitting the trajectory file while using Global Arrays. Under otherwise identical conditions as in the previous section we now observed near ideal scaling behavior with efficiencies above 0.9 (Figure 5b and 5c) without any straggler tasks (Figure 5e). Although the reason why in our case Global Arrays appeared to be more efficient than direct MPI-based communication remains unclear, these results showed that contention for file access clearly impacted performance. By removing the contention, near ideal scaling could be achieved.

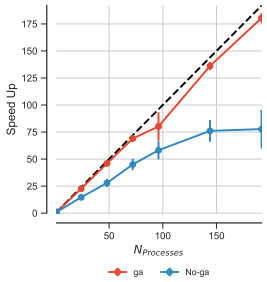
Further considerations for splitting trajectories. The subfiling approach appears promising but it requires preprocessing of trajectory files and additional storage space for the segments. We benchmarked the necessary time for splitting



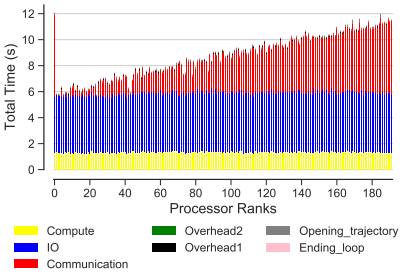
(a) Scaling for different components



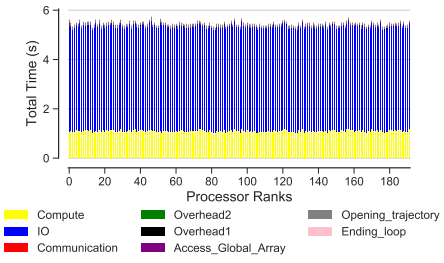
(b) Scaling total



(c) Speed-up



(d) Time comparison on different parts of the calculations per MPI rank without Global Arrays



(e) Time comparison on different parts of the calculations per MPI rank using Global Arrays

Figure 5: Comparison on the performance of the RMSD task ($\overline{t_{\mathrm{comp}}}/\overline{t_{\mathrm{I/O}}} \approx 0.3$) on *SDSC Comet* when the trajectories are split. For communicating the results, either Global Arrays ("ga") or MPI (no Global Arrays, "No-ga") was used. In the case of Global Arrays, all ranks updates the Global Arrays (`ga_put()`) and rank 0 accessed the whole RMSD array through the global memory address (`ga_get()`). Five repeats were performed to collect statistics. (a) Compute and I/O scaling versus number of processes (b) Total time scaling versus number of processes (c) Speed-up (a-c) The error bars show standard deviation with respect to mean. (d-e) Compute t_{comp}, IO $t_{\mathrm{I/O}}$, communication t_{comm}, access to the whole Global Arrays by rank 0, $t_{\mathrm{Access\_Global\_Array}}$, ending the for loop $t_{\mathrm{end\_loop}}$, opening the trajectory $t_{\mathrm{opening\_trajectory}}$, and overheads $t_{\mathrm{overhead1}}$, $t_{\mathrm{overhead2}}$ per MPI rank (see Table 3 for the definitions). When Global Arrays was not used, the performance was decreased due to the non-uniform communication time across different ranks. However, with Global Arrays communication time was significantly reduced and scaling was close to ideal. **Note:** In serial, there is no communication.

the trajectory using different number of MPI processes (Table 5); in general the operation scales well although performance can fluctuate, as seen for the case on 6 nodes. These preprocessing times were not included in the estimates because we are focusing on better understanding the principle causes of stragglers and we wanted to make the results directly comparable to the results of the previous sections. Nevertheless, from an end user perspective, preprocessing of

N_{seg}	N_{p}	nodes	time (s)	S	E
24	24	1	89.9	1.0	1.0
48	48	2	46.8	1.9	0.96
72	72	3	33.7	2.7	0.89
96	96	4	25.1	3.6	0.89
144	144	6	43.7	2.1	0.34
192	192	8	13.5	6.7	0.83

Table 5: The wall-clock time spent for writing N_{seg} trajectory segments using N_{p} processes using MPI on *SDSC Comet*. One set of runs was performed for the timings. Scaling S and efficiency E are relative to the 1 node case (24 MPI processes).

trajectories can be integrated in workflows (especially as the data in Table 5 indicate good scaling) and the preprocessing time can be quickly amortized if the trajectories are analyzed repeatedly.

Often trajectories from MD simulations on HPC machines are produced and kept in small chunks that would need to be concatenated to form a trajectory but that might be useful for the subfiling approach. However, it might not be feasible to have exactly one trajectory segment per MPI rank. In Appendix C we investigated if existing functionality in MDAnalysis that can create virtual trajectories from trajectory segments could benefit from the subfiling approach. We found some improvements in performance but discovered limitations in the design that first have to be overcome before equivalent performance can be reached.

6.4.2. MPI-based Parallel HDF5

Another approach we examined to improve I/O scaling was MPI-based Parallel HDF5. We converted our XTC trajectory file into a simple custom HDF5 format so that we could test the performance of parallel I/O with the HDF5 file format. The code for this file format conversion can be found in the GitHub repository. The time it took to convert our XTC file with $2,512,200$ frames into HDF5 format was about 5,400 s on a local workstation with network file system (NFS).

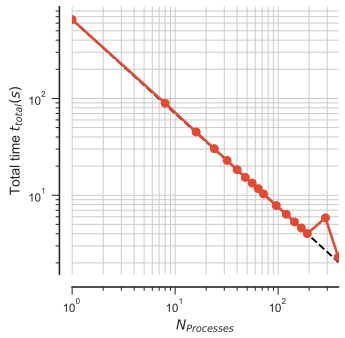
We ran our benchmark on up to 16 nodes (384 cores) and we observed near ideal scaling behavior with parallel efficiencies above 0.8 up to 8 nodes

(Figure A.12a and Figures 6a and 6b) with no straggler tasks (Figure 6d). The trajectory reading I/O ($t_{\mathrm{I/O}}$) was the dominant contribution, followed by compute (t_{comp}), but because both contributions scaled well, overall scaling performance remained good, even for 384 cores. We observed a low-performing outlier for 12 nodes (288 cores) with slower I/O than typical but did not further investigate. Importantly, the trajectory opening cost remained negligible (in the millisecond range) and the cost for MPI communication also remained small (below 0.1 s, even for 16 nodes). Overall, parallel MPI with HDF5 appears to be a robust approach to obtain good scaling, even for I/O-bound tasks.

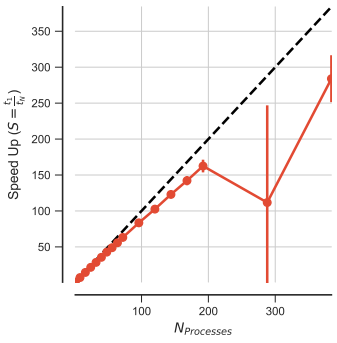
6.5. Likely Causes of Stragglers

The data indicated that increases in the duration of both MPI communication and trajectory file access lead to large variability in the run time of individual MPI processes and ultimately poor scaling performance beyond a single node. A discussion of likely causes for stragglers begins with the observation that opening and reading a single trajectory file from multiple MPI processes appeared to be at the center of the problem.

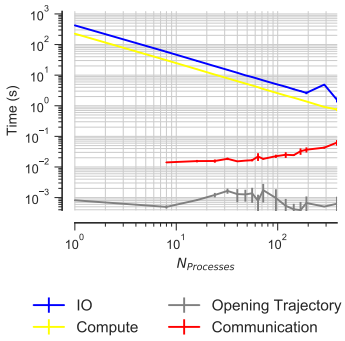
In MDAnalysis, individual trajectory frames are loaded into memory, which ensures that even systems with tens of millions of atoms can be efficiently analyzed on resources with moderate RAM sizes. The test trajectory (file size 30 GB) had $2,512,200$ frames in total so each frame was about 0.011 MB in size. With $t_{\mathrm{I/O}} \approx 0.3$ ms per frame, the data were ingested at a rate of about 40 MB/s for a single process. For 24 MPI ranks (one *SDSC Comet* node), the aggregated reading rate would have been about 1 GB/s. Although, such a value seemed to be well within the available bandwidth of 56 Gb/s of the InfiniBand network interface that served the Lustre file system, in practice the aggregated I/O bandwidth for reading from a shared file was actually fairly small which was expected based on the studies in [61]. Furthermore, in our study the default Lustre stripe size value was 1 MB, i.e., the amount of contiguous data stored on a single Lustre object storage target (OST). Each I/O request read a single Lustre stripe but because the I/O size (0.011 MB) was smaller than the



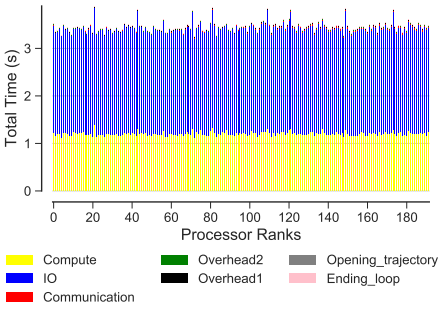
(a) Scaling total



(b) Speed-up



(c) Scaling for different components



(d) Time comparison on different parts of the calculations per MPI rank

Figure 6: Performance of the RMSD task with MPI-based parallel HDF5 ($\overline{t_{\mathrm{comp}}}/\overline{t_{\mathrm{I/O}}} \approx 0.3$) on SDSC Comet. Data are read from the file system from a shared HDF5 file format instead of XTC format (Independent I/O) and results are communicated back to rank 0 (communications included). Five repeats were performed to collect statistics. (a-c) The error bars show standard deviation with respect to mean. (d) Compute t_{comp}, IO $t_{\mathrm{I/O}}$, communication t_{comm}, ending the for loop $t_{\mathrm{end\_loop}}$, opening the trajectory $t_{\mathrm{opening\_trajectory}}$, and overheads $t_{\mathrm{overhead1}}$, $t_{\mathrm{overhead2}}$ per MPI rank (See Table 3 for the definition). This is typical data from one run of the 5 repeats. No straggler is observed. **Note:** In serial, there is no communication. I am reporting the slowest rank per timing, and that is averaged over all repeats.

stripe size, many of these I/O requests were likely just accessing the same stripe on the same OST but nevertheless had to acquire a new reading lock for each request.

The reason for this behavior is related to ensuring POSIX consistency that relates to POSIX I/O API and POSIX I/O semantics, which can have adverse effects on scalability and performance. Parallel file systems like Lustre imple-

ment sophisticated distributed locking mechanisms to ensure consistency. For example, locking mechanisms ensures that a node can not read from a file or part of a file while it might be being modified by another node. In fact, when the application I/O is not designed in a way to avoid scenarios where multiple nodes are fighting over locks for overlapping extents, Lustre can suffer from scalability limitations [62]. Continuously keeping metadata updated in order to have fully consistent reads and writes (POSIX metadata management), requires writing a new value for the file's last-accessed time (POSIX atime) every time a file is read, imposing a significant burden on parallel file system [63].

It was observed that contention for the interconnect between OSTs and compute nodes due to MPI communication may lead to variable performance in I/O measurements [64]. Conversely, our data suggest that single-shared-file I/O on Lustre can negatively affect MPI communication as well, even at moderate numbers (tens to hundreds) of concurrent requests, similar to recent network simulations that predicted interference between MPI and I/O traffic [58]. This work indicated that MPI traffic (inter-process communication) can be affected by increasing I/O, and in particular, a few MPI processes were always delayed by 1-2 orders of magnitude more than the median time. Thus, suggesting that our observed stragglers with large variance in the MPI communication might be due to interference with the I/O requests on the same network.

7. Reproducibility and Performance Comparison on Different Clusters

In this section we compare the performance of the RMSD task on different HPC resources (Table 1) to examine the robustness of the methods we used for our performance study and to ensure that the results are general and independent from the specific HPC system. In Appendix A, we already demonstrated that stragglers occur on *PSC Bridges* and *LSU SuperMIC* in a manner similar to the one observed on *SDSC Comet*. We performed additional comparisons for several cases discussed previously, namely (1) splitting the trajectories with

blocking collective communications in MPI, (2) splitting the trajectories with Global Arrays for communications, and (3) MPI-based parallel HDF5.

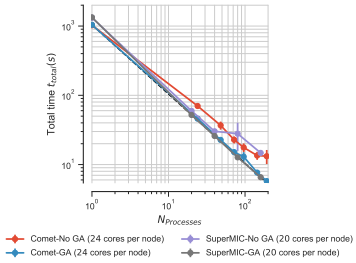
7.1. Splitting the Trajectories

Figure 7 shows the strong scaling of the RMSD task on different HPC resources. Splitting the trajectories with Global Arrays for communication resulted in very good scaling performance on *SuperMIC*, similar to the results obtained on *Comet*. The results with MPI blocking collective communication (instead of Global Arrays) were also comparable between the two clusters, with scaling far from ideal due to the communication cost (see section 6.4.1 and Figures 5d and A.13). Overal, the scaling of the RMSD task is better on *SuperMIC* than on *Comet* and the performance gap increased with increasing core numbers. The results on *SuperMIC* confirmed the conclusion obtained on *Comet* that at least in this case Global Arrays performed better than MPI blocking collective communication.

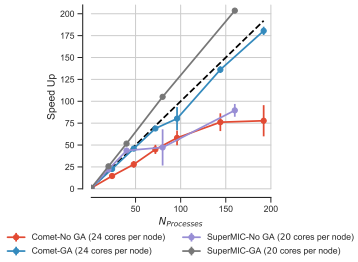
7.2. MPI-based Parallel HDF5

Figure 8 shows the scaling on *SDSC Comet*, *LSU SuperMIC*, and *PSC Bridges* using MPI-based parallel HDF5. Performance on *Comet* and *SuperMIC* was very good with near ideal linear strong scaling. The performance on *Bridges* was sensitive to how many cores per node were used. Using all 28 cores in a node resulted in poor performance but decreasing the number of cores per node and equally distributing processes over nodes improved the scaling (Figure 8), mainly by reducing variation in the I/O times.

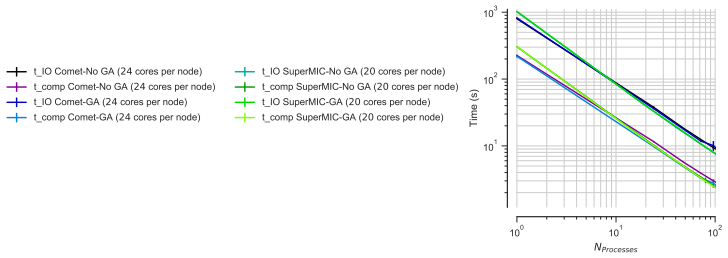
The main difference between the runs on *Bridges* and *Comet*/*SuperMIC* appeared to be the variance in $t_{\mathrm{I/O}}$ (Figure 8c). The I/O time distribution was fairly small and uniform across all ranks on *Comet* and *SuperMIC* (Figures 9b and 6d). However, on *Bridges* the I/O time was on average about two and a half times larger and the I/O time distribution was also more variable across different ranks (Figure 9a).



(a) Scaling total



(b) Speed-up



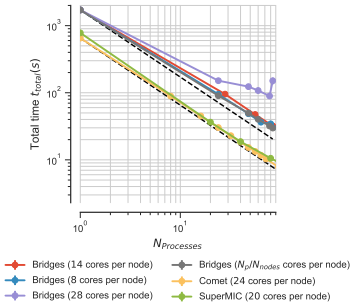
(c) Scaling of the t_{comp} and $t_{\mathrm{I/O}}$ of the RMSD task with MPI when the trajectories are split using Global Arrays and without Global Arrays.

Figure 7: Comparison of the performance of the RMSD task with MPI ($t_{\mathrm{comp}}/t_{\mathrm{I/O}} \approx 0.3$) when the trajectories are split using Global Arrays and without Global Arrays (using MPI for communications) across different clusters (*SDSC Comet*, *LSU SuperMIC*). In case of Global Arrays, all ranks update the Global Arrays (`ga_put()`) and rank 0 accesses the whole RMSD array through the global memory address (`ga_get()`). Five repeats were performed to collect statistics. The error bars show standard deviation with respect to mean.

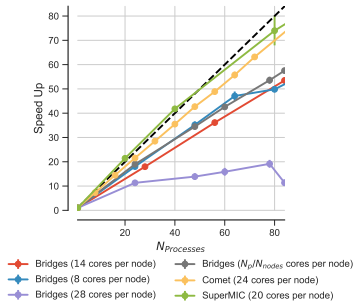
7.3. Comparison of Compute and I/O Scaling Across Different Clusters

A full comparison of compute and I/O scaling across different clusters for different test cases and algorithms is shown in Table 6. For MPI-based parallel HDF5, both the compute and I/O time on *Bridges* were consistenly larger than their corresponding values on *Comet* and *SuperMIC*. For example, with one core the corresponding compute and I/O time were $t_{\mathrm{comp}} = 387$ s, $t_{\mathrm{I/O}} = 1318$ s versus 225 s, 423 s on *Comet* and 273 s, 503 s on *SuperMIC*. This performance difference became larger with increasing core numbers. When the trajectories were split and Global Arrays was used for communication both *Comet* and *SuperMIC* showed similar performance.

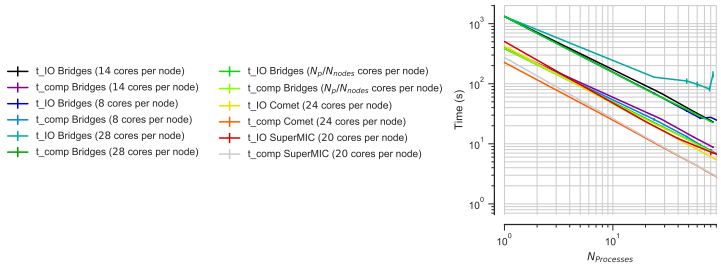
Overall, the results from *Comet* and *SuperMIC* are consistent with each



(a) Scaling total



(b) Speed-up



(c) Scaling of the t_{comp} and $t_{\mathrm{I/O}}$ of the RMSD task with MPI-based parallel HDF5 (Independent I/O)

Figure 8: Comparison of the performance of the RMSD task with MPI ($t_{\mathrm{comp}}/t_{\mathrm{I/O}} \approx 0.3$) across different clusters (*SDSC Comet*, *PSC Bridges*, *LSU SuperMIC*). Data are read from a shared HDF5 file format instead of XTC format (independent I/O) and results are communicated back to rank 0. $N_{\mathrm{P}}/N_{\mathrm{nodes}}$ means that number of processes used for the task is equal on all compute nodes. Five repeats were performed to collect statistics. The error bars show standard deviation with respect to mean.

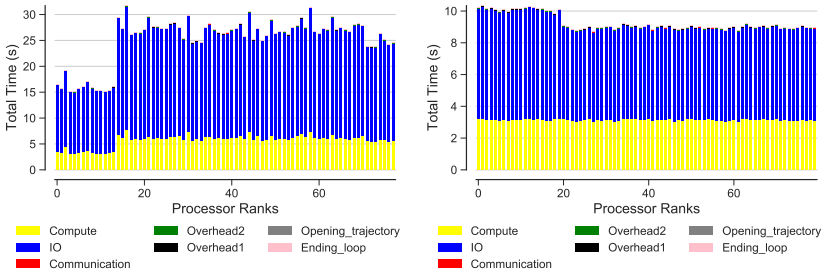
other. Performance on *Bridges* seemed sensitive to the exact allocation of cores on each node but nevertheless the approaches that decreased the occurence of stragglers on *Comet* and *SuperMIC* also improved performace on *Bridges*. Thus, the findings described in the previous sections are valid for a range of different HPC clusters with Lustre file systems.

8. Performance Tuning (Rules of Thumb)

Based on the performance statistics presented in the previous sections we formulate heuristics to improve the performance of parallel analysis of MD trajectories with *MDAnalysis* on HPC resources. However, the proposed guidelines should also be more generally applicable to trajectory data analysis with other

Cluster	Gather	File Access	Time	$N_{Processes}$							
				Serial	Comet: 24 Bridges: 24 SuperMIC: 20	Comet: 48 Bridges: 48 SuperMIC: 40	Comet: 72 Bridges: 60 SuperMIC: 80	Comet: 96 Bridges: 78	Comet: 144 Bridges: 84 SuperMIC: 160	Comet: 192	Comet: 384 SuperMIC: 320
Comet	MPI	Single	$t_{I/O}$ t_{comp}	791 ± 5.22 225 ± 5.4	49 ± 3.45 11 ± 0.75	29 ± 1.3 6 ± 0.35	26 ± 9.19 4 ± 0.48	–	–	–	–
Bridges	MPI	Single	$t_{I/O}$ t_{comp}	770 ± 10.8 221 ± 3.9	38 ± 0.84 11 ± 0.43	33 ± 19.4 6 ± 0.32	15 ± 1.6 4 ± 0.18	–	–	–	–
SuperMIC	MPI	Single	$t_{I/O}$ t_{comp}	1014.51 ± 2.94 303.85 ± 2.3	48.08 ± 0.35 14.56 ± 0.14	24.5 ± 0.79 7.4 ± 0.25	12 ± 0.31 3.7 ± 0.12	–	6.24 ± 0.38 1.8 ± 0.04	–	–
Comet	GA	Single	$t_{I/O}$ t_{comp}	820 ± 18.49 219 ± 9.8	41 ± 8.99 10 ± 0.3	23 ± 4.14 5 ± 0.48	15 ± 2.06 3 ± 0.54	–	–	–	–
Comet	MPI	Splitting	$t_{I/O}$ t_{comp}	799 ± 5.22 225 ± 5.4	37 ± 1.22 11 ± 0.31	18 ± 0.18 5 ± 0.07	12 ± 0.14 3 ± 0.04	9 ± 0.3 3 ± 0.11	6 ± 0.66 2 ± 0.23	4 ± 0.23 1 ± 0.07	–
SuperMIC	MPI	Splitting	$t_{I/O}$ t_{comp}	1013.75 ± 2.8 304.26 ± 2.55	39.99 ± 0.36 12.41 ± 0.22	19.18 ± 0.25 5.99 ± 0.09	9.61 ± 0.28 3.08 ± 0.13	–	4.83 ± 0.06 1.5 ± 0.01	–	–
Comet	GA	Splitting	$t_{I/O}$ t_{comp}	820 ± 18.5 219 ± 9.5	36 ± 0.78 9 ± 0.22	17 ± 0.3 4 ± 0.07	11 ± 0.23 3 ± 0.04	10 ± 1.7 2 ± 0.4	5 ± 0.14 1 ± 0.05	4 ± 0.07 1 ± 0.02	–
SuperMIC	GA	Splitting	$t_{I/O}$ t_{comp}	1027.62 ± 10.32 305.78 ± 3.47	39.62 ± 0.2 12.16 ± 0.1	19.66 ± 0.1 6.01 ± 0.007	9.57 ± 0.1 2.97 ± 0.1	–	4.86 ± 0.05 1.51 ± 0.03	–	–
Comet	MPI	PHDF5	$t_{I/O}$ t_{comp}	423 ± 5.88 225 ± 6.55	19 ± 0.3 10 ± 0.12	9 ± 0.13 5 ± 0.1	6 ± 0.06 3 ± 0.04	5 ± 0.12 2 ± 0.05	3 ± 0.2 1 ± 0.04	3 ± 0.25 1 ± 0.03	1.57 ± 0.29 0.76 ± 0.09
Bridges	MPI	PHDF5	$t_{I/O}$ t_{comp}	1318.87 ± 10.42 387.8 ± 5.51	67.93 ± 0.52 21.97 ± 0.38	37.37 ± 0.2 12.12 ± 0.34	30.35 ± 0.15 9.79 ± 0.24	24.16 ± 0.89 7.72 ± 0.03	22.5 ± 0.17 7.18 ± 0.08	–	–
SuperMIC	MPI	PHDF5	$t_{I/O}$ t_{comp}	503.69 ± 2.57 273.54 ± 4.7	12.96 ± 0.06 23.44 ± 0.29	6.46 ± 0.02 12.22 ± 0.43	3.2 ± 0.01 7.3 ± 0.85	–	1.64 ± 0.01 4.59 ± 0.96	–	0.82 ± 0.004 1.55 ± 0.009

Table 6: Comparison of the compute and I/O scaling for different test cases and number of processes. Five repeats were performed to collect statistics. The mean value and the standard deviation with respect to mean are reported for each case.



(a) Bridges (b) SuperMIC

Figure 9: Examples of timing per MPI rank for RMSD task with MPI-based parallel HDF5 ($t_{\text{comp}}/t_{\text{I/O}} \approx 0.3$) on (a) *PSC Bridges* and (b) *LSU SuperMIC*. Five repeats were performed to collect statistics and these were typical data from one run of the 5 repeats. Compute t_{comp}, I/O $t_{\text{I/O}}$, communication t_{comm}, ending the for loop $t_{\text{end\_loop}}$, opening the trajectory $t_{\text{opening\_trajectory}}$, and overheads $t_{\text{overhead1}}$, $t_{\text{overhead2}}$ per MPI rank (See Table 3 for the definition).

Python-based libraries. To achieve near ideal scaling we suggest the following step by step guidelines:

Heuristic 1 Calculate compute to I/O ($\overline{t_{\text{comp}}}/\overline{t_{\text{I/O}}}$) ratio and compute to communication ($\overline{t_{\text{comp}}}/\overline{t_{\text{comm}}}$) ratio. The $\overline{t_{\text{comp}}}/\overline{t_{\text{I/O}}}$ ratio determines whether the task is compute bound ($\frac{\overline{t_{\text{comp}}}}{\overline{t_{\text{I/O}}}} \gg 1$) or IO bound ($\frac{\overline{t_{\text{comp}}}}{\overline{t_{\text{I/O}}}} \ll 1$). The $\overline{t_{\text{comp}}}/\overline{t_{\text{comm}}}$ ratio determines whether the task is communication bound ($\frac{\overline{t_{\text{comp}}}}{\overline{t_{\text{comm}}}} \ll 1$) or compute bound ($\frac{\overline{t_{\text{comp}}}}{\overline{t_{\text{I/O}}}} \gg 1$).

As discussed in Section 6.2, for I/O bound problems the interference between MPI communication and I/O traffic can be problematic [46, 58, 65] and the performance of the task will be affected by the straggler tasks which delay job completion time.

Heuristic 2 For $\frac{\overline{t_{\text{comp}}}}{\overline{t_{\text{I/O}}}} \geqslant 1$, single-shared-file I/O can be used safely and will not affect performance depending on how large is compute with respect to I/O. Therefore, one needs to consider the following cases:

Heuristic 2.1 If $\frac{\overline{t_{\text{comp}}}}{\overline{t_{\text{comm}}}} \gg 1$, the task is compute bound and will scale well as shown in Figure 2.

Heuristic 2.2 If $\frac{\overline{t_{\text{comp}}}}{\overline{t_{\text{comm}}}} \ll 1$, one might consider using Global Arrays to achieve near ideal scaling behavior (Section 6.3). Application of Global

Arrays replaces the message-passing interface with a distributed shared array where its blocks will be updated by the associated rank in the communication domain and collection of all data can be performed efficiently (Algorithm 2).

Heuristic 3 For $\frac{\overline{t_{\mathrm{comp}}}}{\overline{t_{\mathrm{I/O}}}} \leqslant 1$ the task is I/O bound and single-shared-file I/O should be avoided to remove unnecessary metadata operations. Therefore, one needs to take the following steps:

Heuristic 3.1 If there is access to HDF5 format the recommended way will be to use MPI-based Parallel HDF5 (Section 6.4.2).

Heuristic 3.2 If the trajectory file is not in HDF5 format then one might prefer to split into as many trajectory segments as the number of processes. Splitting the trajectories can be easily performed in parallel as opposed to converting the XTC file to HDF5, which is computationally more expensive. MD trajectories are often re-analyzed and therefore incorporating trajectory conversion into the beginning of standard workflows for MD simulations could improve the performance of such workflows. Alternatively, it may be beneficial to keep the trajectories in smaller chunks, e.g., when running simulations on HPC resources using Gromacs [66, 67], users can run their simulations with the "-noappend" option so that output trajectories will be automatically stored in small chunks.

Heuristic 3.3 In case of $\frac{\overline{t_{\mathrm{comp}}}}{\overline{t_{\mathrm{comm}}}} \ll 1$, appropriate parallel implementation along with *Global Arrays* should be used on the trajectory segments (Section 6.4.1) to achieve near ideal scaling.

9. Conclusion

We analyzed the strong scaling performance of a typical task when analysing MD trajectories, the calculation of the time series of the RMSD of a protein,

with the widely used Python-based *MDAnalysis* library. The task was parallelized with MPI by having each MPI process analyze a contiguous segment of the trajectory. This approach did not scale beyond a single node because straggler processes exhibited large upward variations in runtime. Stragglers were primarily caused by either excessive MPI communication costs or excessive time to open the single shared trajectory file whereas both the computation and the ingestion of data exhibited close to ideal strong scaling behavior. Stragglers were less prevalent for compute-bound workloads (i.e., $\overline{t_{\mathrm{comp}}}/\overline{t_{\mathrm{I/O}}} \gg 1$ and to a lesser degree $\overline{t_{\mathrm{comp}}}/\overline{t_{\mathrm{comm}}} \gg 1$), suggesting that file I/O was responsible for poor MPI communication. In particular, artifically removing all I/O substantially improved performance of the communication step and thus brought overall performance close to ideal. By performing benchmarks on three different XSEDE supercomputers we showed that our results were independent from the specifics of the hardware and local environment. Our results hint at the possibility that stragglers might be due to the competition between MPI messages and the Lustre file system on the shared InfiniBand interconnect, which would be consistent with other similar observations [46, 65] and theoretical predictions [58]. In this model, I/O interferes much less often with communication for a sufficiently large per-frame compute workload than for an I/O bound task. This analysis also suggested ways to improve performance by improving I/O or improving communication.

When we used the Global Arrays toolkit instead of the message-passing interface for communication, then the communication cost was significantly reduced and there were no delayed task due to communication. Despite these improvements, overall performance remained less than ideal because initial opening of the shared trajectory file became the performance bottleneck which is due to the POSIX consistency requirements. We were able to eliminate this I/O problem by subfiling (splitting the shared trajectory file) and achieved nearly ideal scaling up to 192 cores (8 nodes on *SDSC Comet*, Table 6). Alternatively, we used MPI-based parallel I/O through HDF5 and MPI for communications with comparable performance up to 384 cores (16 nodes on *SDSC Comet*, Table 6).

The latter approach appears to be a promising way forward as it directly builds on very widely used technology (HDF5 and MPI-IO) and echoes the experience of the wider HPC community that parallel file I/O is necessary for efficient data handling. The biomolecular simulation community suffers from a large number of trajectory file formats with very few being based on HDF5, with the exception of the H5MD format [68] and the MDTraj HDF5 format [69]. Our work suggests that HDF5-based formats should be seriously considered as the default for MD simulations if users want to make efficient use of their HPC systems for analysis.

These strategies also provide guidelines for parallel analysis on data generated from MD simulations. We showed that it is feasible to run an I/O bound analysis task on HPC resources and achieve good scaling behavior up to 16 nodes with an almost 300-fold speed-up compared to serial execution. Although we focused on the *MDAnalysis* library, similar strategies are likely to be more generally applicable and useful to the wider biomolecular simulation community.

Acknowledgements We are grateful to Sarp Oral for insightful comments on this manuscript. Funding: This work was supported by the National Science Foundation [grant numbers 1443054,1440677]. Computational resources were provided by NSF XRAC awards TG-MCB090174 and TG-MCB130177.

References

1. Borhani DW, Shaw DE. The future of molecular dynamics simulations in drug discovery. J Comput Aided Mol Des 2012;26(1):15–26. doi:10.1007/s10822-011-9517-y.
2. Dror RO, Dirks RM, Grossman JP, Xu H, Shaw DE. Biomolecular simulation: a computational microscope for molecular biology. Annu Rev Biophys 2012;41:429–52. doi:10.1146/annurev-biophys-042910-155245.
3. Orozco M. A theoretical view of protein dynamics. Chem Soc Rev 2014;43:5051–66. doi:10.1039/C3CS60474H.
4. Perilla JR, Goh BC, Cassidy CK, Liu B, Bernardi RC, Rudack T, et al. Molecular dynamics simulations of large macromolecular complexes. Current Opinion in Structural Biology 2015;31(0):64 – 74. URL: http://www.sciencedirect.com/science/article/pii/S0959440X15000342. doi:http://dx.doi.org/10.1016/j.sbi.2015.03.007.

5. Bottaro S, Lindorff-Larsen K. Biophysical experiments and biomolecular simulations: A perfect match? Science 2018;361(6400):355–60. doi:10.1126/science.aat4010.

6. Mura C, McAnany CE. An introduction to biomolecular simulations and docking. Molecular Simulation 2014;40(10-11):732–64. URL: http://dx.doi.org/10.1080/08927022.2014.935372. doi:10.1080/08927022.2014.935372. arXiv:http://dx.doi.org/10.1080/08927022.2014.935372.

7. Cheatham T, Roe D. The impact of heterogeneous computing on workflows for biomolecular simulation and analysis. Computing in Science Engineering 2015;17(2):30–9. doi:10.1109/MCSE.2015.7.

8. Kneller GR, Keiner V, Kneller M, Schiller M. nmoldyn: A program package for a neutron scattering oriented analysis of molecular dynamics simulations. Computer Physics Communications 1995;91(1):191 – 214. URL: http://www.sciencedirect.com/science/article/pii/001046559500048K. doi:http://dx.doi.org/10.1016/0010-4655(95)00048-K.

9. Hinsen K, Pellegrini E, Stachura S, Kneller GR. nmoldyn 3: Using task farming for a parallel spectroscopy-oriented analysis of molecular dynamics simulations. Journal of Computational Chemistry 2012;33(25):2043–8. URL: http://dx.doi.org/10.1002/jcc.23035. doi:10.1002/jcc.23035.

10. Humphrey W, Dalke A, Schulten K. VMD – Visual Molecular Dynamics. J Mol Graph 1996;14:33–8. URL: http://www.ks.uiuc.edu/Research/vmd/.

11. Hinsen K. The molecular modeling toolkit: a new approach to molecular simulations. Journal of Computational Chemistry 2000;21(2):79–85.

12. Grant BJ, Rodrigues APC, ElSawy KM, McCammon JA, Caves LSD. Bio3d: an r package for the comparative analysis of protein structures. Bioinformatics 2006;22(21):2695–6. doi:10.1093/bioinformatics/btl461.

13. Tu T, Rendleman CA, Borhani DW, Dror RO, Gullingsrud J, Jensen MO, et al. A scalable parallel framework for analyzing terascale molecular dynamics simulation trajectories. In: 2008 SC - International Conference for High Performance Computing, Networking, Storage and Analysis. 2008, p. 1–12. doi:10.1109/SC.2008.5214715.

14. Romo TD, Grossfield A. LOOS: An extensible platform for the structural analysis of simulations. In: 31st Annual International Conference of the IEEE EMBS. Minneapolis, Minnesota, USA: IEEE; 2009, p. 2332–5.

15. Romo TD, Leioatts N, Grossfield A. Lightweight object oriented structure analysis: Tools for building tools to analyze molecular dynamics simulations. Journal of Computational Chemistry 2014;35(32):2305–18. URL: http://dx.doi.org/10.1002/jcc.23753. doi:10.1002/jcc.23753.

16. Michaud-Agrawal N, Denning EJ, Woolf TB, Beckstein O. MDAnalysis: A toolkit for the analysis of molecular dynamics simulations. J Comp Chem 2011;32:2319–27.

doi:10.1002/jcc.21787.

17. Gowers RJ, Linke M, Barnoud J, Reddy TJE, Melo MN, Seyler SL, et al. MDAnalysis: A Python package for the rapid analysis of molecular dynamics simulations. In: Benthall S, Rostrup S, editors. Proceedings of the 15th Python in Science Conference. Austin, TX: SciPy; 2016, p. 102 –9. URL: http://mdanalysis.org.

18. Roe DR, Thomas E. Cheatham I. Ptraj and cpptraj: Software for processing and analysis of molecular dynamics trajectory data. Journal of Chemical Theory and Computation 2013;9(7):3084–95. URL: http://dx.doi.org/10.1021/ct400341p. doi:10.1021/ct400341p. arXiv:http://dx.doi.org/10.1021/ct400341p; pMID: 26583988.

19. McGibbon RT, Beauchamp KA, Harrigan MP, Klein C, Swails JM, Hernández CX, et al. Mdtraj: A modern open library for the analysis of molecular dynamics trajectories. Biophysical Journal 2015;109(8):1528 –32. URL: //www.sciencedirect.com/science/article/pii/S0006349515008267. doi:http://dx.doi.org/10.1016/j.bpj.2015.08.015.

20. Yesylevskyy SO. Pteros 2.0: Evolution of the fast parallel molecular analysis library for c++ and python. Journal of Computational Chemistry 2015;36(19):1480–8. URL: http://dx.doi.org/10.1002/jcc.23943. doi:10.1002/jcc.23943.

21. Doerr S, Harvey MJ, Noé F, De Fabritiis G. HTMD: High-throughput molecular dynamics for molecular discovery. Journal of Chemical Theory and Computation 2016;12(4):1845–52. URL: http://dx.doi.org/10.1021/acs.jctc.6b00049. doi:10.1021/acs.jctc.6b00049.

22. Khoshlessan M, Paraskevakos I, Jha S, Beckstein O. Parallel analysis in MDAnalysis using the Dask parallel computing library. In: Katy Huff , David Lippa , Dillon Niederhut , Pacer M, editors. Proceedings of the 16th Python in Science Conference. Austin, TX: SciPy; 2017, p. 64–72. doi:10.25080/shinma-7f4c6e7-00a.

23. Paraskevakos I, Luckow A, Khoshlessan M, Chantzialexiou G, Cheatham TE, Beckstein O, et al. Task-parallel analysis of molecular dynamics trajectories. In Proceedings of 47th International Conference on Parallel Processing; University of Oregon, Eugene, Oregon, USA: ICPP; 2018,.

24. Liu P, Agrafiotis DK, Theobald DL. Fast determination of the optimal rotational matrix for macromolecular superpositions. J Comput Chem 2010;31(7):1561–3. doi:10.1002/jcc.21439.

25. Rocklin M. Dask: Parallel computation with blocked algorithms and task scheduling. In: Proceedings of the 14th Python in Science Conference. 130–136; 2015,URL: https://github.com/dask/dask.

26. Dalcín LD, Paz RR, Kler PA, Cosimo A. Parallel distributed computing using python. Advances in Water Resources 2011;34(9):1124 –39. doi:10.1016/j.advwatres.2011.

04.013; new Computational Methods and Software Tools.

27. Dalcín L, Paz R, Storti M. MPI for python. Journal of Parallel and Distributed Computing 2005;65(9):1108 –15. doi:10.1016/j.jpdc.2005.03.010.
28. Straggler Root-Cause and Impact Analysis for Massive-scale Virtualized Cloud Datacenters. ISSN 1939-1374; https://doi.org/10.1109/TSC.2016.2611578; 2016.
29. Phan TD. Energy-efficient straggler mitigation for big data applications on the clouds. Ph.D. thesis; École normale supérieure de Renne; 2017.
30. DAILY JA. Gain: Distributed array computation with python. Master's thesis; School of Electrical Engineering and Computer Science, WASHINGTON STATE UNIVERSITY; 2009.
31. Nieplocha J, Palmer B, Tipparaju V, Krishnan M, Trease H, Aprà E. Advances, applications and performance of the global arrays shared memory programming toolkit. The International Journal of High Performance Computing Applications 2006;20(2):203–31.
32. Dean J, Ghemawat S. Mapreduce: Simplified data processing on large clusters. In: OSDI'04 Sixth Symposium on Operating System Design and Implementation. 2004, p. pp. 137–150.
33. Kyong J, Jeon J, Lim SS. Improving scalability of apache spark-based scale-up server through docker container-based partitioning. In: Proceedings of the 6th International Conference on Software and Computer Applications - ICSCA '17. New York, USA: ACM Press. ISBN 9781450348577; 2017, p. 176–80. URL: http://dl.acm.org/citation.cfm?doid=3056662.3056686. doi:10.1145/3056662.3056686.
34. Ousterhout K. Architecting for Performance Clarity in Data Analytics Frameworks 2017;URL: http://www2.eecs.berkeley.edu/Pubs/TechRpts/2017/EECS-2017-158.html.
35. Gittens A, Devarakonda A, Racah E, Ringenburg M, Gerhardt L, Kottalam J, et al. Matrix factorizations at scale: A comparison of scientific data analytics in spark and c+mpi using three case studies. In: IEEE International Conference on Big Data (Big Data). ISBN 978-1-4673-9005-7; 2016, p. 204–13. URL: http://ieeexplore.ieee.org/document/7840606/. doi:10.1109/BigData.2016.7840606.
36. Schmidt E, DeMichillie G, Perry F, Akidau T, Halperin. D. Large-scale data analysis at cloud scale. In: Symposium on Frontiers in Big Data. 2016,.
37. Qi Chen CL, Xiao Z. Improving mapreduce performance using smart speculative execution strategy. In: IEEE Transactions on Computers; vol. 63 of *DOI: 10.1109/TC.2013.15*. IEEE; 2014, p. 954 –67.
38. Xie B, Chase J, Dillow D, Drokin O, Klasky S, Oral S, et al. Characterizing output bottlenecks in a supercomputer. In: Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis. SC '12; Los Alamitos, CA, USA: IEEE Computer Society Press. ISBN 978-1-4673-0804-5; 2012, p. 8:1–8:11.

URL: http://dl.acm.org/citation.cfm?id=2388996.2389007.

39. Yang H, Liu X, Chen S, Lei Z, Du H, Zhu C. Improving Spark performance with MPTE in heterogeneous environments. In: 2016 International Conference on Audio, Language and Image Processing (ICALIP). IEEE. ISBN 978-1-5090-0654-0; 2016, p. 28–33. URL: http://ieeexplore.ieee.org/document/7846627/. doi:10.1109/ICALIP.2016.7846627.
40. Rosen J. Fine-grained micro-tasks for mapreduce skew-handling; 2012.
41. Kwon Y, Balazinska M, Howe B, Rolia J. Skewtune: Mitigating skew in mapreduce applications, pages 25-36. In: SIGMOD'12. DOI: 10.1145/2213836.2213840: SIGMOD '12 Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data; 2012, p. Pages 25–36.
42. Ousterhout K, Rasti R, Ratnasamy S, Shenker S, Chun BG. Making sense of performance in data analytics frameworks. In: NSDI'15 Proceedings of the 12th USENIX Conference on Networked Systems Design and Implementation. ISBN: 978-1-931971-218; 2015, p. Pages 293–307.
43. Abdul-Wahid B, Feng H, Rajan D, Costaouec R, Darve E, Thain D, et al. Awe-wq, fast-forwarding molecular dynamics using the accelerated weighted ensemble. Journal of Chemical Information and Modeling 2014;54:3033–43.
44. Wu G, Song H, Lin D. A scalable parallel framework for microstructure analysis of large-scale molecular dynamics simulations data. Computational Materials Science 2018;144:322–30.
45. Tu T, Rendleman CA, Miller PJ, Sacerdoti F, Dror RO, Shaw DE. Accelerating parallel analysis of scientific simulation data via zazen. In: 8th USENIX Conference on File and Storage Technologies, San Jose, CA, USA. http://www.usenix.org/events/fast10/tech/full_papers/tu.pdf; 2010, p. 129–42.
46. Stone JE, Isralewitz B, Schulten K. Early experiences scaling vmd molecular visualization and analysis jobs on blue waters. In: Proceedings of the 2013 Extreme Scaling Workshop (Xsw 2013). Washington, DC, USA: IEEE Computer Society; 2013, p. 43–50.
47. Shkurtia A, Goni R, Andrio P, Breitmoserd E, Bethuned I, Orozco M, et al. pypcazip: A pca-based toolkit for compression and analysis of molecular simulation data. SoftwareX 2016;5:44–50.
48. Malakar P, Knight C, Munson T, Vishwanath V, Papka ME. Scalable in situ analysis of molecular dynamics simulations. In: ISAV'17 Proceedings of the In Situ Infrastructures on Enabling Extreme-Scale Analysis and Visualization. 2017, p. 1–6.
49. Leach AR. Molecular Modelling. Principles and Applications. Longman; 1996.
50. Theobald DL. Rapid calculation of RMSDs using a quaternion-based characteristic polynomial. Acta Crystallogr A 2005;61(Pt 4):478–80. doi:10.1107/S0108767305015266.

51. Daily J, Vishnu A, Palmer B, van Dam H, Kerbyson D. On the suitability of MPI as a PGAS runtime. In: 2014 21st International Conference on High Performance Computing (HiPC). 2014, p. 1–10. doi:10.1109/HiPC.2014.7116712.

52. Collette A. Python and HDF5. O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.; 2014.

53. Seyler SL, Beckstein O. Sampling of large conformational transitions: Adenylate kinase as a testing ground. Molec Simul 2014;40(10–11):855–77. doi:10.1080/08927022.2014.919497.

54. Seyler S, Beckstein O. Molecular dynamics trajectory for benchmarking MDAnalysis. 2017. URL: https://figshare.com/articles/Molecular_dynamics_trajectory_for_benchmarking_MDAnalysis/5108170. doi:10.6084/m9.figshare.5108170.

55. Shaw DE, Dror RO, Salmon JK, Grossman JP, Mackenzie KM, Bank JA, et al. Millisecond-scale molecular dynamics simulations on anton. In: SC '09: Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis. New York, NY, USA: ACM. ISBN 978-1-60558-744-8; 2009, p. 1–11. doi:10.1145/1654059.1654099.

56. Salomon-Ferrer R, Götz AW, Poole D, Le Grand S, Walker RC. Routine microsecond molecular dynamics simulations with amber on gpus. 2. explicit solvent particle mesh ewald. Journal of Chemical Theory and Computation 2013;9(9):3878–88. URL: http://pubs.acs.org/doi/abs/10.1021/ct400314y. doi:10.1021/ct400314y. arXiv:http://pubs.acs.org/doi/pdf/10.1021/ct400314y.

57. Glaser J, Nguyen TD, Anderson JA, Lui P, Spiga F, Millan JA, et al. Strong scaling of general-purpose molecular dynamics simulations on gpus. Computer Physics Communications 2015;192:97–107. URL: http://www.sciencedirect.com/science/article/pii/S0010465515000867. doi:dx.doi.org/10.1016/j.cpc.2015.02.028.

58. Brown KA, Jain N, Matsuoka S, Schulz M, Bhatele A. Interference between I/O and MPI traffic on fat-tree networks. In: Proceedings of the 47th International Conference on Parallel Processing. ICPP 2018; New York, NY, USA: ACM. ISBN 978-1-4503-6510-9; 2018, p. 7:1–7:10. URL: http://doi.acm.org/10.1145/3225058.3225144. doi:10.1145/3225058.3225144.

59. Choudhary A, keng Liao W, Gao K, Nisar A, Ross R, Thakur R, et al. Scalable i/o and analytics. Journal of Physics: Conference Series 2009;180(012048).

60. Son SW, Sehrish S, keng Liao W, Oldfield R, Choudhary A. Reducing i/o variability using dynamic i/o path characterization in petascale storage systems. Journal of Supercomputing 2017;73(5):pp 2069–2097.

61. Latham R, Carns P. Thinking about hpc io and hpc storage. In: ATPESC. 2016,.

62. Lin KW, Chou J, Byna S, and KW. Optimizing fast query performance on Lustre file system. In: SSDBM Proceedings of the 25th International Conference on Scientific and Statistical Database Management. Article No. 29; 2013,.

63. https://www.nextplatform.com/2017/09/11/whats-bad-posix-io/, what is so bad about posix i/o? ????

64. Mache J, Lo V, Garg S. The impact of spatial layout of jobs on I/O hotspots in mesh networks. Journal of Parallel and Distributed Computing 2005;65(10):1190 –203. URL: http://www.sciencedirect.com/science/article/pii/S0743731505001048. doi:10.1016/j.jpdc.2005.04.020; design and Performance of Networks for Super-, Cluster-, and Grid-Computing Part I.

65. Brown KA, Jain N, Matsuoka S, Schulz M, Bhatele A. Interference between io and mpi traffic on fat-tree networks. In: Proceedings of the 47th International Conference on Parallel Processing. No. 7 in ICPP 2018; New York, NY, USA: ACM; 2018, p. 7:1–7:10.

66. Lindahl E, Hess B, van der Spoel D. Gromacs 3.0: A package for molecular simulation and trajectory analysis. J Mol Mod 2001;7(8):306–17. URL: http://www.gromacs.org. doi:10.1007/s008940100045.

67. Berendsen HJC, van der Spoel D, van Drunen R. GROMACS: A message-passing parallel molecular dynamics implementation. Comp Phys Comm 1995;91:43–56.

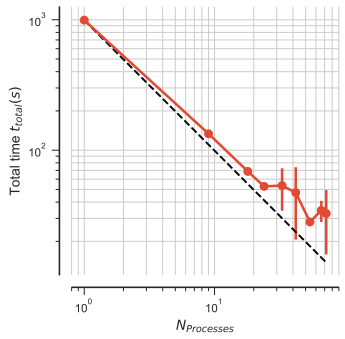
68. de Buyl P, Colberg PH, Höfling F. H5MD: A structured, efficient, and portable file format for molecular data. Computer Physics Communications 2014;185(6):1546 –53. doi:10.1016/j.cpc.2014.01.018.

69. McGibbon RT, Beauchamp KA, Harrigan MP, Klein C, Swails JM, Hernández CX, et al. MDTraj: A modern open library for the analysis of molecular dynamics trajectories. Biophysical Journal 2015;109(8):1528 –32. URL: http://www.sciencedirect.com/science/article/pii/S0006349515008267. doi:10.1016/j.bpj.2015.08.015.

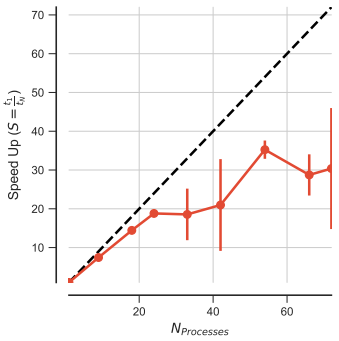
70. Lindahl E, Hess B, van der Spoel D. Gromacs 3.0: A package for molecular simulation and trajectory analysis. J Mol Mod 2001;7(8):306–17. URL: http://www.gromacs.org. doi:10.1007/s008940100045.

Appendix A. Additional Data

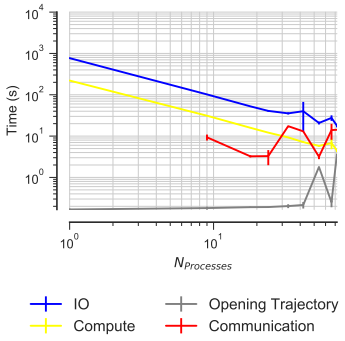
Figure A.10 shows performance of the RMSD task on PSC Bridges.



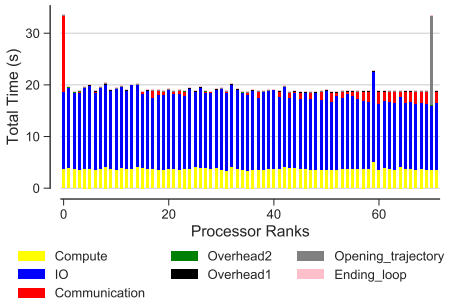
(a) Scaling total (five repeats)



(b) Speed-up (five repeats)



(c) Scaling for different components (five repeats)



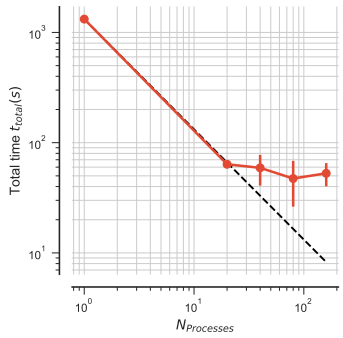
(d) Time comparison on different parts of the calculations per MPI rank (example)

Figure A.10: Performance of the RMSD task with MPI which is I/O-bound $\overline{t_{\text{comp}}}/\overline{t_{\text{I/O}}} \approx 0.3$ on PSC Bridges. Results are communicated back to rank 0 (communications included). Five independent repeats were performed to collect statistics. (a-c) The error bars show standard deviation with respect to mean. (d) Compute t_{comp}, IO $t_{\text{I/O}}$, communication t_{comm}, ending the for loop $t_{\text{end\_loop}}$, opening the trajectory $t_{\text{opening\_trajectory}}$, and overheads $t_{\text{overhead1}}$, $t_{\text{overhead2}}$ per MPI rank (See Table 3 for the definition). These are data from one run of the five repeats. MPI ranks 0 and 70 are stragglers. **Note:** In serial, there is no communication.

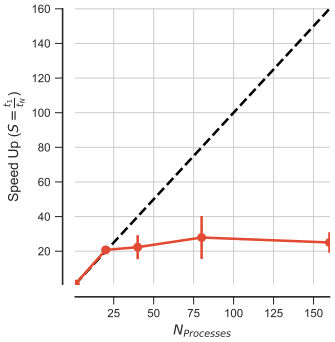
Figure A.11 shows performance of the RMSD task on SuperMIC.

Figure A.12 shows comparison of the parallel efficiency of the RMSD task between different test cases on SDSC Comet, PSC Bridges, and LSU SuperMIC.

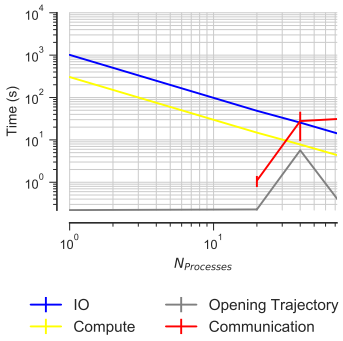
Figure A.13 shows how RMSD task scales with the increase in the number of cores when the trajectories are split using GA and without GA on SuperMIC.



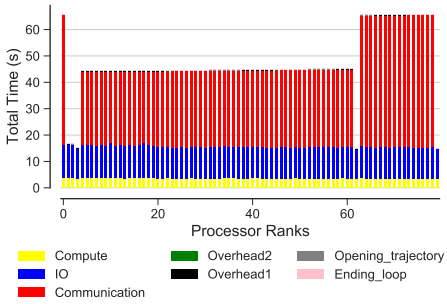
(a) Scaling total (five repeats)



(b) Speed-up (five repeats)



(c) Scaling for different components (five repeats)



(d) Time comparison on different parts of the calculations per MPI rank (example)

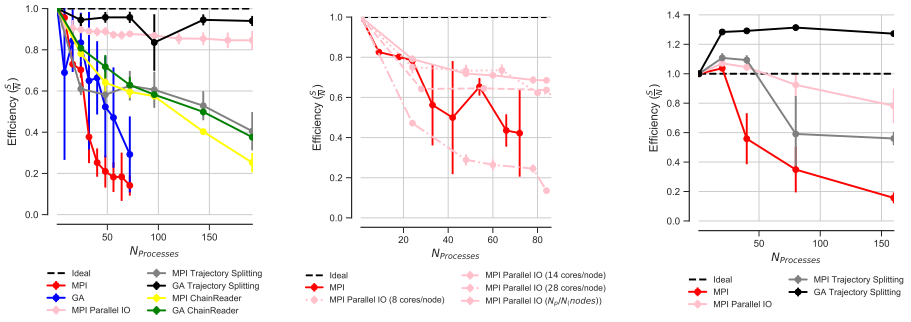
Figure A.11: Performance of the RMSD task with MPI which is I/O-bound $\overline{t_{\text{comp}}}/\overline{t_{\text{I/O}}} \approx 0.3$ on SuperMIC. Results are communicated back to rank 0 (communications included). Five independent repeats were performed to collect statistics. (a-c) The error bars show standard deviation with respect to mean. (d) Compute t_{comp}, IO $t_{\text{I/O}}$, communication t_{comm}, ending the for loop $t_{\text{end\_loop}}$, opening the trajectory $t_{\text{opening\_trajectory}}$, and overheads $t_{\text{overhead1}}$, $t_{\text{overhead2}}$ per MPI rank (See Table 3 for the definition). These are data from one run of the five repeats. MPI ranks 0 and 70 are stragglers. **Note:** In serial, there is no communication.

Appendix B. Effect of $t_{\text{comp}}/t_{\text{comm}}$ on Performance

In addition to the compute to I/O ratio discussed in Section 6.2 we define another performance parameter called the compute to communication ratio $t_{\text{comp}}/t_{\text{comm}}$. In Section 6.4, we overcame the I/O effect by splitting the trajectory, but scaling remained far from ideal when MPI communication was used (instead of Global Arrays). This is because the task remained communication bound (Figure 5), i.e,

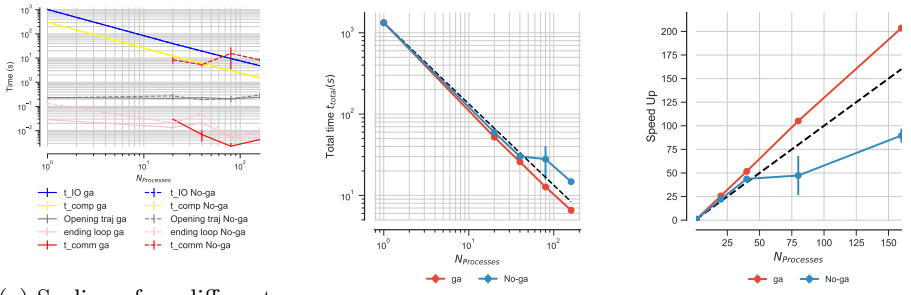
$$\frac{t_{\text{comp}}}{t_{\text{comm}}} \ll 1.$$

Figure B.14 shows the relationship of performance with $\overline{t_{\text{comp}}}/\overline{t_{\text{comm}}}$ ratio. When the



(a) Comet (b) Bridges (c) SuperMIC

Figure A.12: Comparison of the parallel efficiency between different test cases on (a) SDSC Comet, (b) Bridges, and (c) SuperMIC. Five repeats were performed to collect statistics and error bars show standard deviation with respect to mean.



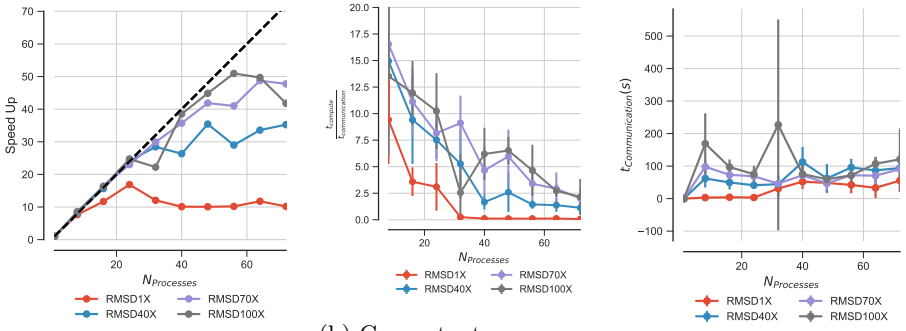
(a) Scaling for different components (b) Scaling total (c) Speed-up

Figure A.13: Comparison on the performance of the RMSD task with MPI when the trajectories are split using global array and without global array ($\overline{t_{\text{comp}}}/\overline{t_{\text{I/O}}} \approx 0.3$) on SuperMIC. In case of global array, all ranks update the global array (ga_{put}) and rank 0 accesses the whole RMSD array through the global memory address (ga_{get}). Five repeats were performed to collect statistics. (a) Compute & I/O scaling versus number of processes (b) Total time scaling versus number of processes (c) Speed-up (a-c) The error bars show standard deviation with respect to mean.

$\overline{t_{\text{comp}}}/\overline{t_{\text{comm}}}$ ratio is higher (Figure B.14b), performance is better (Figure B.14a) even if communication time is larger (Figure B.14c). Although, we still observed stragglers due to communication at larger $\overline{t_{\text{comp}}}/\overline{t_{\text{comm}}}$ ratios (70× RMSD and 100× RMSD), their effect on performance remained modest because the overall performance was dominated by the compute load. Evidently, if overall performance is dominated by a component such as compute that scales well, then performance problems with components such as communication will be masked and overall acceptable performance can still be achieved (Figures B.14a and B.14b).

Communication is usually not problematic within one node because of the shared memory environment (for less than 24 processes (single compute node on *SDSC Comet*) the scaling is good and $\frac{t_{\text{comp}}}{t_{\text{comm}}} \gg 1$ for all RMSD loads) (Figures B.14a and B.14b). However, beyond a

single compute node, scaling appears to get better as the $\overline{t_{\mathrm{comp}}}/\overline{t_{\mathrm{comm}}}$ ratio increases and communication overhead becomes less dominant (Figures B.14a and B.14b, 24-72 cores represent multiple compute nodes on *SDSC Comet*).



(a) Speed-Up (b) Compute to communication ratio (c) Communication time

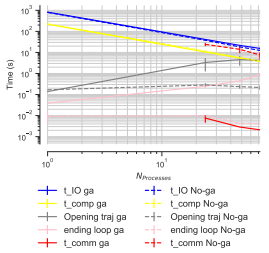
Figure B.14: (a) Change in compute to communication ratio with number of processes for different RMSD workload on SDSC Comet. (b) Comparison of communication time for different RMSD workload on SDSC Comet. Five repeats were performed to collect statistics and error bars show standard deviation with respect to mean.

Appendix C. Performance of the ChainReader for split trajectories

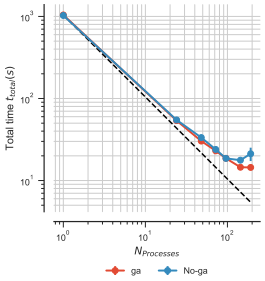
In Section 6.4.1 we showed how splitting the trajectories would help to overcome I/O and improve scaling. However, the number of trajectories may not necessarily be equal to the number of processes. For example, trajectories from MD simulations on HPC machines are often kept in small chunks that need to be concatenated to form a trajectory that can be analyzed with the typical tools. For the subfiling (splitting trajectories) such chunks might be useful but making sure that the number of processes is equal to the number of trajectory files will not always be feasible for the typical users. *MDAnalysis* can transparently represent multiple trajectories as one virtual trajectory using the "ChainReader". This feature is already very convenient for serial analysis when trajectories are maintained as chunks. In the current implementation of ChainReader, each process opens all the trajectories but I/O will only happen from a specific block of the trajectory specific to that process only.

We wanted to test if the ChainReader would benefit from the gains measured for the subfiling approach. Specifically, we measured if the MPI-parallelized RMSD task (with N_{p} ranks) would benefit if the trajectory was split into $N_{\mathrm{seg}} = N_{\mathrm{p}}$ trajectory segments, corresponding to an ideal scenario.

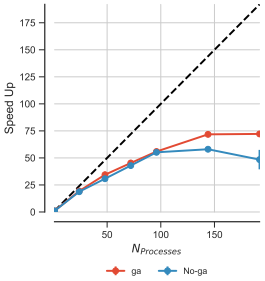
In order to perform our experiments we had to work around a design problem in the XTC format reader in MDAnalysis; the Gromacs XTC format [70] is a lossy-compression,



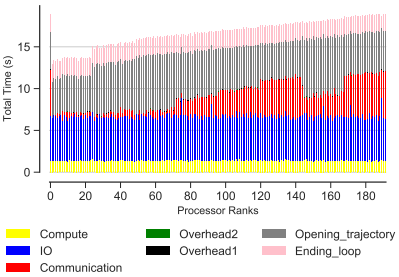
(a) Scaling for different components



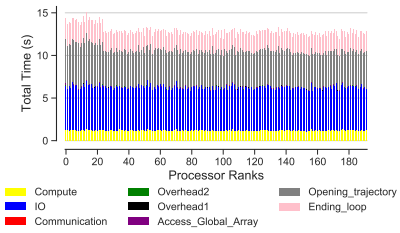
(b) Scaling total



(c) Speed-up



(d) Time comparison on different parts of the calculations per MPI rank using ChainReader without Global Arrays



(e) Time comparison on different parts of the calculations per MPI rank using ChainReader using Global Arrays

Figure C.15: Comparison od the performance of the MDAnalysis ChainReader for the RMSD task ($\overline{t_{\mathrm{comp}}}/\overline{t_{\mathrm{I/O}}} \approx 0.3$) with MPI on *SDSC Comet* when the trajectories are split; for the communication Global Arrays ("ga") or MPI (without Global Arrays, "no-ga") were used. In case of Global Arrays, all ranks update the Global Arrays (`ga_put()`) and rank 0 accesses the whole RMSD array through the global memory address (`ga_get()`). Five repeats were performed to collect statistics. (a) Compute & I/O scaling versus number of processes (b) Total time scaling versus number of processes (c) Speed-up (a-c) The error bars show standard deviation with respect to mean. (d-e) Compute t_{comp}, IO $t_{\mathrm{I/O}}$, communication t_{comm}, access to the whole Global Arrays by rank 0 $t_{\mathrm{Access\_Global\_Array}}$, ending the for loop $t_{\mathrm{end\_loop}}$, opening the trajectory $t_{\mathrm{opening\_trajectory}}$, and overheads $t_{\mathrm{overhead1}}$, $t_{\mathrm{overhead2}}$ per MPI rank (See Table 3 for the definition). When Global Arrays was not used, the performance was affected due to the non-uniform communication time across different ranks. However, with Global Arrays communication time was significantly reduced and scaling was close to ideal. In addition, time for ending the for loop $t_{\mathrm{end\_loop}}$ and opening the trajectory $t_{\mathrm{opening\_trajectory}}$ became a bottleneck as opposed to our calculation without ChainReader. We obtained the results for the ChainReader with an unreleased patched version of MDAnalysis that avoids a race condition. Note: In serial, there is no communication.

XDR-based file format that was never designed for parallel file access. The XTCReader stores persistent offsets to disk [17] in order to enable efficient access to random frames because the compressed XTC format itself does not support fast random seeking. These offsets will be generated automatically the first time the trajectory is opened and are stored in hidden $*.xtc\_offsets.npz$ files. The advantage of these persistent offset files is that after opening the

trajectory for the first time, opening the same file will be very fast afterward. However, stored offsets can get out of sync with the trajectory they refer to. To prevent the use of stale offset data, trajectory file data are (number of atoms, size of the file and last modification time) are also stored for validation. If any of these parameters change the offsets are recalculated. If the XTC changes but the offset file is not updated then the offset file can be detected as invalid. With ChainReader in parallel, each process opens all the trajectories because each process builds its own MDAnalysis.Universe data structure; if an invalid offset file is detected (perhaps because of wrong file modification timestamps across nodes), several processes might want to recalculate these parameters and rebuild the offset file, which can lead to race conditions. In order to avoid the race condition, we removed this check from MDAnalysis, but this comes at the price of not checking the validity of the offset files at all; future versions of MDAnalysis will lift this limitation. We obtained the results for the ChainReader with a *patched version* of *MDAnalysis* that eliminates the race condition by assuming that XTC index files are present and valid.

Figure C.15 shows the results for performance of the ChainReader for the RMSD task using GA and without GA (i.e., using MPI for communications). As shown in Figure C.15c the cases with GA and without GA scaled up to 144 and 92 cores respectively. The scaling was not close to ideal as opposed to what we achieved in Section 6.4.1 when each MPI process was assigned its own trajectory segment. However, in this case I/O and compute time were scaling very well (Figure C.15a as compared to Figure 5a). The time for ending the `for` loop $t_{\text{end\_loop}}$ (which includes the time for closing the trajectory file) and opening the trajectory $t_{\text{opening\_trajectory}}$ appeared to be the performance bottleneck as opposed to the results shown in Section 6.4.1 (i.e. Figures 5d and 5e).

Although we did not further investigate the deeper cause for the reduced performance of the ChainReader, we speculate that the primary problem is related to each MPI rank having to open all trajectory files in their ChainReader instance even though they will only read from a small subset. For N_{p} ranks and N_{seg} file segments, in total, $N_p N_{\text{seg}}$ file opening/closing operations have to be performed. Each server that is part of a Lustre filesystem can only handle a limited number of I/O requests (read, write, stat, open, close, etc.) per second. An excessive number of such requests, from one or more users and one or more jobs, can lead to contention for storage resources. For $N_{\text{p}} = N_{\text{seg}} = 100$, the Lustre file system has to perform 10,000 of these operations almost simultaneously, which might degrade performance.